SPRINGER REFERENCE

Bruce Mizrach
Brian Dangerfield
Section Editors

Robert A. Meyers
Editor-in-Chief

Complex Systems in Finance and Econometrics

Selected entries from the Encyclopedia of Complexity and Systems Science

Springer

Complex Systems in Finance and Econometrics

This book consists of selections from the
Encyclopedia of Complexity and Systems Science
edited by Robert A. Meyers,
published by Springer New York in 2009.

Robert A. Meyers (Ed.)

Complex Systems in Finance and Econometrics

With 282 Figures and 54 Tables

Springer

Robert A. Meyers, Ph. D.
Editor-in-Chief
RAMTECH LIMITED
122 Escalle Lane
Larkspur, CA 94939
USA
robert.meyers@ramtechlimited.org

Library of Congress Control Number: 2010938142

ISBN: 978-1-4419-7701-4
This publication is available also as:
Print publication under ISBN: 978-1-4419-7700-7 and
Print and electronic bundle under ISBN 978-1-4419-7702-1

© 2011 SpringerScience+Buisiness Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

This book consists of selections from the *Encyclopedia of Complexity and Systems Science* edited by Robert A. Meyers, published by Springer New York in 2009.

springer.com

Printed on acid free paper

Preface

Complex Systems in Finance and Econometrics is an authoritative reference to the basic tools and concepts of complexity and systems theory as applied to an understanding of complex, financial-based business and social systems. Fractals, nonlinear time series modeling, cellular automata, game theory, network theory and statistical physics are among the tools and techniques that are used for predicting, monitoring, evaluating, managing, and decision-making in a wide range of fields from health care, poverty alleviation, and energy and the environment, to manufacturing and quality assurance, model building, organizational learning. and macro and microeconomics. In the last of these areas, market bubbles and crashes, foreign exchange, and bond markets are addressed.

Sixty-nine of the world's leading experts present 49 articles for an audience of advanced undergraduate and graduate students, professors, and professionals in all of these fields. Each article was selected and peer reviewed by one of the Section Editors of the *Encyclopedia of Complexity and Systems Science* with advice and consultation provided by our Board Members and Editor-in-Chief. This level of coordination assures that the reader can have a level of confidence in the relevance and accuracy of the information far exceeding that generally found on the World Wide Web or any print publication. Accessiblilty is also a priority and for this reason each article includes a glossary of important terms and a concise definition of the subject. The primary Section Editors for this project were Bruce Mizrach and Brian Dangerfield, while Andrej Nowak, Cristina Marchetti, Marilda Sotomayor, Daniel ben-avraham and Schlomo Havlin recruited and reviewed several of the articles. An alphabetical list of the 49 articles and the authors is presented on pages XV through XVII, and the articles are also organized by section on pages VII to VIII. A summary, perspective and roadmap for the articles on Finance and Econometrics can be found on pages 290 to 292, and for System Dynamics on pages 853 to 855.

Complex systems are systems that comprise many interacting parts with the ability to generate a new quality of collective behavior through self-organization, e.g. the spontaneous formation of temporal, spatial or functional structures. They are therefore adaptive as they evolve and may contain self-driving feedback loops. Thus, complex systems are much more than a sum of their parts. Complex systems are often characterized as having extreme sensitivity to initial conditions as well as emergent behavior that are not readily predictable or even completely deterministic. One conclusion is that a reductionist (bottom-up) approach is often an incomplete description of a phenomenon. This recognition, that the collective behavior of the whole system cannot be simply inferred from the understanding of the behavior of the individual components, has led to many new concepts and sophisticated mathematical and modeling tools for application to financial-based business and social systems.

Acknowledgements

I wish to thank Springer management David Packer, Lydia Mueller and Sylvia Blago for selecting this project for publication and providing valuable advice and counsel.

Robert A. Meyers
Editor in Chief
Larkspur, California
August 2010

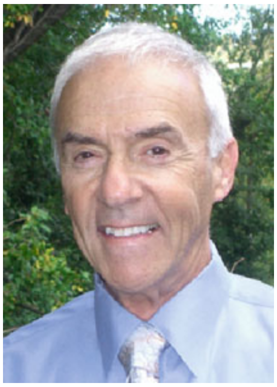
Sections

Finance and Econometrics,
Section Editor: Bruce Mizrach

System Dynamics,
Section Editor: Brian Dangerfield

Additional Economics Approaches,
Section Editors: Daniel ben-Avraham, Filippo Castiglione, Shlomo Havlin,
M. Cristina Marchetti, Andrzej Nowak, and Marilda Sotomayor

About the Editor-in-Chief



Robert A. Meyers

President: RAMTECH Limited
Manager, Chemical Process Technology, TRW Inc.
Post-doctoral Fellow: California Institute of Technology
Ph. D. Chemistry, University of California at Los Angeles
B. A., Chemistry, California State University, San Diego

Biography

Dr. Meyers has worked with more than 25 Nobel laureates during his career.

Research

Dr. Meyers was Manager of Chemical Technology at TRW (now Northrop Grumman) in Redondo Beach, CA and is now President of RAMTECH Limited. He is co-inventor of the Gravimelt process for desulfurization and demineralization of coal for air pollution and water pollution control. Dr. Meyers is the inventor of and was project manager for the DOE-sponsored Magnetohydrodynamics Seed Regeneration Project which has resulted in the construction and successful operation of a pilot plant for production of potassium formate, a chemical utilized for plasma electricity generation and air pollution control. Dr. Meyers managed the pilot-scale DoE project for determining the hydrodynamics of synthetic fuels. He is a co-inventor of several thermo-oxidative stable polymers which have achieved commercial success as the GE PEI, Upjohn Polyimides and Rhone-Polenc bismaleimide resins. He has also managed projects for photochemistry, chemical lasers, flue gas scrubbing, oil shale analysis and refining, petroleum analysis and refining, global change measurement from space satellites, analysis and mitigation (carbon dioxide and ozone), hydrometallurgical refining, soil and hazardous waste remediation, novel polymers synthesis, modeling of the economics of space transportation systems, space rigidizable structures and chemiluminescence-based devices.

He is a senior member of the American Institute of Chemical Engineers, member of the American Physical Society, member of the American Chemical Society and serves on the UCLA Chemistry Department Advisory Board. He was a member of the joint USA-Russia working group on air pollution control and the EPA-sponsored Waste Reduction Institute for Scientists and Engineers.

Dr. Meyers has more than 20 patents and 50 technical papers. He has published in primary literature journals including Science and the Journal of the American Chemical Society, and is listed in Who's Who in America and Who's Who in the World. Dr. Meyers' scientific achievements have been reviewed in feature articles in the popular press in publications such as The New York Times Science Supplement and The Wall Street Journal as well as more specialized publications such as Chemical Engineering and Coal Age. A public service film was produced by the Environmental Protection Agency of Dr. Meyers' chemical desulfurization invention for air pollution control.

Scientific Books

Dr. Meyers is the author or Editor-in-Chief of 12 technical books one of which won the Association of American Publishers Award as the best book in technology and engineering.

Encyclopedias

Dr. Meyers conceived and has served as Editor-in-Chief of the Academic Press (now Elsevier) Encyclopedia of Physical Science and Technology. This is an 18-volume publication of 780 twenty-page articles written to an audience of university students and practicing professionals. This encyclopedia, first published in 1987, was very successful, and because of this, was revised and reissued in 1992 as a second edition. The Third Edition was published in 2001 and is now online. Dr. Meyers has completed two editions of the Encyclopedia of Molecular Cell Biology and Molecular Medicine for Wiley VCH publishers (1995 and 2004). These cover molecular and cellular level genetics, biochemistry, pharmacology, diseases and structure determination as well as cell biology. His eight-volume Encyclopedia of Environmental Analysis and Remediation was published in 1998 by John Wiley & Sons and his 15-volume Encyclopedia of Analytical Chemistry was published in 2000, also by John Wiley & Sons, all of which are available on-line.

Editorial Board Members

Thomas C. Schelling
Distinguished University Professor, University of Maryland, Emeritus
2005 Nobel Prize in Economics for understanding of conflict and cooperation through game-theory analysis
Current interests include: applications of complexity in economics and political science.

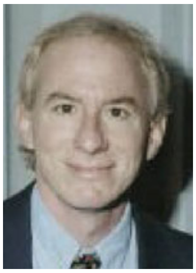
Professor Benoit B. Mandelbrot
Sterling Professor Emeritus of Mathematical Sciences at Yale University
1993 Wolf Prize for Physics and the
2003 Japan Prize for Science and Technology
Current interests include: seeking a measure of order in physical, mathematical or social phenomena that are characterized by abundant data but wild variability.

Stephen Wolfram
Founder and CEO, Wolfram Research
Creator, Mathematica®
Author, *A New Kind of Science*

Richard E. Stearns
1993 Turing Award for foundations
of computational complexity
Current interests include: computational complexity, automata theory, analysis of algorithms, and game theory.

Section Editors

Finance and Econometrics



BRUCE MIZRACH
Professor
Department of Economics
Rutgers University

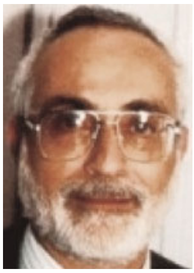
System Dynamics

BRIAN DANGERFIELD
Professor of Systems Modelling & Executive Editor
System Dynamics Review Centre for OR & Applied
Statistics Salford Business School
Faculty of Business, Law & the Built Environment
University of Salford

Contributing Section Editors

DANIEL BEN-AVRAHAM
Professor
Department of Physics
Clarkson University

FILIPPO CASTIGLIONE
Research Scientist
Institute for Computing Applications (IAC) “M. Picone”
National Research Council (CNR), Italy



Shlomo Havlin
Professor
Department of Physics
Bar Ilan University

Andrzej Nowak
Director of the Center for Complex Systems
University of Warsaw
Assistant Professor, Psychology Department
Florida Atlantic Universityh

M. Cristina Marchetti
William R. Kenan, Jr. Professor of Physics
Physics Department
Syracuse University

Marilda Sotomayor
Professor
Department of Economics
University of São Paulo, Brazil
Department of Economics
Brown University, Providence

Table of Contents

Contributors

ACKERMANN, FRAN
University of Strathclyde
Glasgow
UK

ANDERSEN, DAVID F.
University at Albany
Albany
USA

ANDERSEN, TORBEN G.
Northwestern University
Evanston
USA
NBER
Cambridge
USA
CREATES
Aarhus
Denmark

BENZONI, LUCA
Federal Reserve Bank of Chicago
Chicago
USA

DANGERFIELD, BRIAN
University of Salford
Salford
UK

DIKS, CEES
University of Amsterdam
Amsterdam
The Netherlands

EDEN, COLIN
University of Strathclyde
Glasgow
UK

ESCANCIANO, JUAN-CARLOS
Indiana University
Bloomington
USA

ESCRIBANO, ALVARO
Universidad Carlos III de Madrid
Madrid
Spain

FORD, ANDREW
Washington State University, Pullman
Washington
USA

GALLEGATI, MAURO
Università Politecnica delle Marche
Ancona
Italy

GEORGANTZAS, NICHOLAS C.
Fordham University Business Schools
New York
USA

GONZÁLEZ-RIVERA, GLORIA
University of California
Riverside
USA

GROESSER, STEFAN
University of St. Gallen
St. Gallen
Switzerland

HAAS, MARKUS
University of Munich
Munich
Germany

HAFNER, CHRISTIAN M.
Université catholique de Louvain
Louvain-la-Neuve
Belgium

HIRSCH, GARY
Independent Consultant
Wayland
USA

HOMER, JACK
Independent Consultant
Voorhees
USA

HOWICK, SUSAN
University of Strathclyde
Glasgow
UK

KAMPMANN, CHRISTIAN ERIK
Copenhagen Business School
Copenhagen
Denmark

KAMSTRA, MARK J.
York University
Toronto
Canada

KEILIS-BOROK, VLADIMIR
University of California
Los Angeles
USA
Russian Academy of Science
Moscow
Russia

KOIJEN, RALPH S. J.
Tilburg University
Tilburg
The Netherlands

KORENOK, OLEG
VCU School of Business
Richmond
USA

KRAMER, LISA A.
University of Toronto
Toronto
Canada

LEE, TAE-HWY
University of California
Riverside
USA

LEVY, MOSHE
The Hebrew University
Jerusalem
Israel

LICHTMAN, ALLAN
American University
Washington D.C.
USA

LYNEIS, JAMES M.
Worcester Polytechnic Institute
Worcester
USA

MAANI, KAMBIZ
The University of Queensland
Brisbane
Australia

MACDONALD, RODERICK
University at Albany
Albany
USA

MAIER, FRANK H.
International University in Germany
Bruchsal
Germany

MANZAN, SEBASTIANO
Baruch College CUNY
New York
USA

MARKELLOS, RAPHAEL N.
Loughborough University
Loughborough
UK
Athens University of Economics and Business
Athens
Greece

MILLER, CHRISTIAN S.
Board of Governors of the Federal Reserve System
Washington DC
USA

MILLING, PETER M.
Mannheim University
Mannheim
Germany

MILLS, TERENCE C.
Loughborough University
Loughborough
UK

MIZRACH, BRUCE
Rutgers University
New Brunswick
USA

MORLEY, JAMES
Washington University
St. Louis
USA

Moss, Scott
Manchester Metropolitan University Business School
Manchester
UK

Neely, Christopher J.
Federal Reserve Bank of St. Louis
St. Louis
USA

Olaya, Camilo
Universidad de Los Andes
Bogotá
Colombia

Oliva, Rogelio
Texas A&M University
College Station
USA

Osler, Carol
Brandeis University
Waltham
USA

Petkova, Ralitsa
Texas A&M University
College Station
USA

Piger, Jeremy
University of Oregon
Eugene
USA

Pigorsch, Christian
University of Bonn
Bonn
Germany

Piwowar, Michael S.
Securities Litigation and Consulting Group, Inc.
Fairfax
USA

Radzicki, Michael J.
Worcester Polytechnic Institute
Worcester
USA

Richardson, George P.
University at Albany, State University of New York
Albany
USA

Rich, Eliot
University at Albany
Albany
USA

Richiardi, Matteo G.
Università Politecnica delle Marche
Ancona
Italy
Collegio Carlo Alberto – LABORatorio R. Revelli
Moncalieri
Italy

Roehner, Bertrand M.
University of Paris 7
Paris
France

Rouwette, Etiënne A. J. A.
Radboud University
Nijmegen
The Netherlands

Saeed, Khalid
Worcester Polytechnic Institute
Worcester
USA

Schwaninger, Markus
University of St. Gallen
St. Gallen
Switzerland

Shintani, Mototsugu
Vanderbilt University
Nashville
USA

Soloviev, Alexandre
Russian Academy of Science
Moscow
Russia
Abdus Salam International Centre for Theoretical Physics
Trieste
Italy

Takayasu, Hideki
Sony Computer Science Laboratories Inc
Tokyo
Japan

Takayasu, Misako
Tokyo Institute of Technology
Tokyo
Japan

TRIVEDI, PRAVIN K.
Indiana University
Bloomington
USA

VAN NIEUWERBURGH, STIJN
New York University
New York
USA

VEGA, CLARA
Board of Governors of the Federal Reserve System
Washington DC
USA

VENNIX, JAC A. M.
Radboud University
Nijmegen
The Netherlands

WILLIAMS, TERRY
Southampton University
Southampton
UK

WOLSTENHOLME, ERIC
South Bank University
London
UK
Symmetric SD
Brighton
UK

WOODERS, MYRNA
Vanderbilt University
Nashville
USA

WOOLDRIDGE, JEFFREY M.
Michigan State University
East Lansing
USA

YAKOVENKO, VICTOR M.
University of Maryland
College Park
USA

Agent Based Computational Economics

Moshe Levy
The Hebrew University, Jerusalem, Israel

Article Outline

Glossary

Agent-based simulation A simulation of a system of multiple interacting agents (sometimes also known as "microscopic simulation"). The "micro" rules governing the actions of the agents are known, and so are their rules of interaction. Starting with some initial conditions, the dynamics of the system are investigated by simulating the state of the system through discrete time steps. This approach can be employed to study general properties of the system, which are not sensitive to the initial conditions, or the dynamics of a specific system with fairly well-known initial conditions, e. g. the impact of the baby boomers' retirement on the US stock market.

Bounded-rationality Most economic models describe agents as being *fully* rational – given the information at their disposal they act in the optimal way which maximizes their objective (or utility) function. This optimization may be technically very complicated, requiring economic, mathematical and statistical sophistication. In contrast, *bounded* rational agents are limited in their ability to optimize. This limitation may be due to limited computational power, errors, or various psychological biases which have been experimentally documented.

Market anomalies Empirically documented phenomena that are difficult to explain within the standard rational representative agent economic framework. Some of these phenomena are the over-reaction and under-reaction of prices to news, the auto-correlation of stock returns, various calendar and day-of-the-week effects, and the excess volatility of stock returns.

Representative agent A standard modeling technique in economics, by which an entire class of agents (e. g. investors) are modeled by a single "representative" agent. If agents are completely homogeneous, it is obvious that the representative agent method is perfectly legitimate. However, when agents are heterogeneous, the representative agent approach can lead to a multitude of problems (see [16]).

Definition of the Subject

Mainstream economic models typically make the assumption that an entire group of agents, e. g. "investors", can be modeled with a single "rational representative agent". While this assumption has proven extremely useful in advancing the science of economics by yielding analytically tractable models, it is clear that the assumption is not realistic: people are different one from the other in their tastes, beliefs, and sophistication, and as many psychological studies have shown, they often deviate from rationality in systematic ways.

Agent Based Computational Economics is a framework allowing economics to expand beyond the realm of the "rational representative agent". By modeling and simulating the behavior of each agent and the interaction among agents, agent based simulation allows us to investigate the dynamics of complex economic systems with many heterogeneous and not necessarily fully rational agents.

The agent based simulation approach allows economists to investigate systems that can not be studied with the conventional methods. Thus, the following key questions can be addressed: How do heterogeneity and systematic deviations from rationality affect markets? Can these elements explain empirically observed phenomena which are considered "anomalies" in the standard economics literature? How robust are the results obtained with the analytical models? By addressing these questions the agent based simulation approach complements the traditional analytical analysis, and is gradually becoming a standard tool in economic analysis.

Introduction

For solving the dynamics of two bodies (e. g. stars) with some initial locations and velocities and some law of attraction (e. g. gravitation) there is a well-known analytical solution. However, for a similar system with three bodies there is no known analytical solution. Of course, this does not mean that physicists can't investigate and predict the behavior of such systems. Knowing the state of the system (i. e. the location, velocity, and acceleration of each body) at time t, allows us to calculate the state of the system an instant later, at time $t + \Delta t$. Thus, starting with the initial conditions we can predict the dynamics of the system by simply simulating the "behavior" of each element in the system over time.

This powerful and fruitful approach, sometimes called "Microscopic Simulation", has been adopted by many other branches of science. Its application in economics is best known as "Agent Based Simulation" or "Agent Based Computation". The advantages of this approach are clear – they allow the researcher to go where no analytical models can go. Yet, despite of the advantages, perhaps surprisingly, the agent based approach was not adopted very quickly by economists. Perhaps the main reason for this is that a particular simulation only describes the dynamics of a system with a particular set of parameters and initial conditions. With other parameters and initial conditions the dynamics may be different. So economists may ask: what is the value of conducting simulations if we get very different results with different parameter values? While in physics the parameters (like the gravitational constant) may be known with great accuracy, in economics the parameters (like the risk-aversion coefficient, or for that matter the entire decision-making rule) are typically estimated with substantial error. This is a strong point. Indeed, we would argue that the "art" of agent based simulations is the ability to understand the general dynamics of the system and to draw general conclusions from a finite number of simulations. Of course, one simulation is sufficient as a counter-example to show that a certain result does not hold, but many more simulations are required in order to convince of an alternative general regularity.

This manuscript is intended as an introduction to agent-based computational economics. An introduction to this field has two goals: (i) to explain and to demonstrate the agent-based methodology in economics, stressing the advantages and disadvantages of this approach relative to the alternative purely analytical methodology, and (ii) to review studies published in this area. The emphasis in this paper will be on the first goal. While Sect. "Some of the Pioneering Studies" does provide a brief review of some of the cornerstone studies in this area, more comprehensive reviews can be found in [19,24,32,39,40], on which part of Sect. "Some of the Pioneering Studies" is based. A comprehensive review of the many papers employing agent based computational models in economics will go far beyond the scope of this article. To achieve goal (i) above, in Sect. "Illustration with the LLS Model" we will focus on one particular model of the stock market in some detail. Section "Summary and Future Directions" concludes with some thoughts about the future of the field.

Some of the Pioneering Studies

Schelling's Segregation Model

Schelling's [36] classical segregation model is one of the earliest models of population dynamics. Schelling's model is not intended as a realistic tool for studying the actual dynamics of specific communities as it ignores economic, real-estate and cultural factors. Rather, the aim of this very simplified model is to explain the emergence of macroscopic single-race neighborhoods even when individuals are not racists. More precisely, Schelling found that the collective effect of neighborhood racial segregation results even from individual behavior that presents only a very mild preference for same-color neighbors. For instance, even the minimal requirement by each individual of having (at least) one neighbor belonging to ones' own race leads to the segregation effect.

The agent based simulation starts with a square mesh, or lattice, (representing a town) which is composed of cells (representing houses). On these cells reside agents which are either "blue" or "green" (the different races). The crucial parameter is the minimal percentage of same-color neighbors that each agent requires. Each agent, in his turn, examines the color of all his neighbors. If the percentage of neighbors belonging to his own group is above the "minimal percentage", the agent does nothing. If the percentage of neighbors of his own color is less then the minimal percentage, the agent moves to the closest unoccupied cell. The agent then examines the color of the neighbors of the new location and acts accordingly (moves if the number of neighbors of his own color is below the minimal percentage and stays there otherwise). This goes on until the agent is finally located at a cite in which the minimal percentage condition holds. After a while, however, it might happen that following the moves of the other agents, the minimal percentage condition ceases to be fulfilled and then the agent starts moving again until he finds an appropriate cell. As mentioned above, the main result is that even for very mild individual preferences for same-color neighbors, after some time the entire system displays a very high level of segregation.

A more modern, developed and sophisticated reincarnation of these ideas is the Sugarscape environment described by Epstein and Axtell [6]. The model considers a population of moving, feeding, pairing, procreating, trading, warring agents and displays various qualitative collective events which their populations incur. By employing agent based simulation one can study the macroscopic results induced by the agents' individual behavior.

The Kim and Markowitz Portfolio Insurers Model

Harry Markowitz is very well known for being one of the founders of modern portfolio theory, a contribution for which he has received the Nobel Prize in economics. It is less well known, however, that Markowitz is also one of the pioneers in employing agent based simulations in economics.

During the October 1987 crash markets all over the globe plummeted by more than 20% within a few days. The surprising fact about this crash is that it appeared to be spontaneous – it was not triggered by any obvious event. Following the 1987 crash researchers started to look for endogenous market features, rather than external forces, as sources of price variation. The Kim-Markowitz [15] model explains the 1987 crash as resulting from investors' "Constant Proportion Portfolio Insurance" (CPPI) policy. Kim and Markowitz proposed that market instabilities arise as a consequence of the individual insurers' efforts to cut their losses by selling once the stock prices are going down.

The Kim Markowitz agent based model involves two groups of individual investors: rebalancers and insurers (CPPI investors). The rebalancers are aiming to keep a constant composition of their portfolio, while the insurers make the appropriate operations to insure that their eventual losses will not exceed a certain fraction of the investment per time period.

The rebalancers act to keep a portfolio structure with (for instance) half of their wealth in cash and half in stocks. If the stock price rises, then the stocks weight in the portfolio will increase and the rebalancers will sell shares until the shares again constitute 50% of the portfolio. If the stock price decreases, then the value of the shares in the portfolio decreases, and the rebalancers will buy shares until the stock again constitutes 50% of the portfolio. Thus, the rebalancers have a stabilizing influence on the market by selling when the market rises and buying when the market falls.

A typical CPPI investor has as his/her main objective not to lose more than (for instance) 25% of his initial wealth during a quarter, which consists of 65 trading days. Thus, he aims to insure that at each cycle 75% of the initial wealth is out of reasonable risk. To this effect, he assumes that the current value of the stock will not fall in one day by more than a factor of 2. The result is that he always keeps in stock twice the difference between the present wealth and 75% of the initial wealth (which he had at the beginning of the 65 days investing period). This determines the amount the CPPI agent is bidding or offering at each stage. Obviously, after a price fall, the amount he wants to keep in stocks will fall and the CPPI investor will sell and further destabilize the market. After an increase in the prices (and personal wealth) the amount the CPPI agent wants to keep in shares will increase: he will buy, and may support a price bubble.

The simulations reveal that even a relatively small fraction of CPPI investors (i. e. less than 50%) is enough to destabilize the market, and crashes and booms are observed. Hence, the claim of Kim and Markowitz that the CPPI policy may be responsible for the 1987 crash is supported by the agent based simulations. Various variants of this model were studied intensively by Egenter, Lux and Stauffer [5] who find that the price time evolution becomes unrealistically periodic for a large number of investors (the periodicity seems related with the fixed 65 days quarter and is significantly diminished if the 65 day period begins on a different date for each investor).

The Arthur, Holland, Lebaron, Palmer and Tayler Stock Market Model

Palmer, Arthur, Holland, Lebaron and Tayler [30] and Arthur, Holland, Lebaron, Palmer and Tayler [3] (AHLPT) construct an agent based simulation model that is focused on the concept of co-evolution. Each investor adapts his/her investment strategy such as to maximally exploit the market dynamics generated by the investment strategies of all others investors. This leads to an ever-evolving market, driven endogenously by the ever-changing strategies of the investors.

The main objective of AHLPT is to prove that market fluctuations may be induced by this endogenous co-evolution, rather than by exogenous events. Moreover, AHLPT study the various regimes of the system: the regime in which rational fundamentalist strategies are dominating vs. the regime in which investors start developing strategies based on technical trading. In the technical trading regime, if some of the investors follow fundamentalist strategies, they may be punished rather than rewarded by the market. AHLPT also study the relation between the various strategies (fundamentals vs. technical) and the volatility properties of the market (clustering, excess volatility, volume-volatility correlations, etc.).

In the first paper quoted above, the authors simulated a single stock and further limited the bid/offer decision to a ternary choice of: i) bid to buy one share, ii) offer to sell one share, or: iii) do nothing. Each agent had a collection of rules which described how he should behave (i, ii or iii) in various market conditions. If the current market conditions were not covered by any of the rules, the default was to do nothing. If more than one rule applied in a certain market condition, the rule to act upon was chosen probabilistically according to the "strengths" of the applicable rules. The "strength" of each rule was determined according to the rule's past performance: rules that "worked" became "stronger". Thus, if a certain rule performed well, it became more likely to be used again.

The price is updated proportionally to the relative excess of offers over demands. In [3], the rules were used to predict future prices. The price prediction was then transformed into a buy/sell order through the use of a Constant Absolute Risk Aversion (CARA) utility function. The use of CARA utility leads to demands which do not depend on the investor's wealth.

The heart of the AHLPT dynamics are the trading rules. In particular, the authors differentiate between "fundamental" rules and "technical" rules, and study their relative strength in various market regimes. For instance, a "fundamental" rule may require a market conditions of the type:

$$\text{dividend/current price} > 0.04$$

in order to be applied. A "technical" rule may be triggered if the market fulfills a condition of the type:

$$\text{current price} > \text{10-period moving-average of past prices.}$$

The rules undergo genetic dynamics: the weakest rules are substituted periodically by copies of the strongest rules and all the rules undergo random mutations (or even versions of "sexual" crossovers: new rules are formed by combining parts from 2 different rules). The genetic dynamics of the trading rules represent investors' learning: new rules represent new trading strategies. Investors examine new strategies, and adopt those which tend to work best. The main results of this model are:

For a Few Agents, a Small Number of Rules, and Small Dividend Changes

- The price converges towards an equilibrium price which is close to the fundamental value.
- Trading volume is low.
- There are no bubbles, crashes or anomalies.
- Agents follow homogeneous simple fundamentalist rules.

For a Large Number of Agents and a Large Number of Rules

- There is no convergence to an equilibrium price, and the dynamics are complex**.**
- The price displays occasional large deviations from the fundamental value (bubbles and crashes).
- Some of these deviations are triggered by the emergence of collectively self-fulfilling agent price-prediction rules.
- The agents become heterogeneous (adopt very different rules).
- Trading volumes fluctuate (large volumes correspond to bubbles and crashes).
- The rules evolve over time to more and more complex patterns, organized in hierarchies (rules, exceptions to rules, exceptions to exceptions, and so on ...).
- The successful rules are time dependent: a rule which is successful at a given time may perform poorly if reintroduced after many cycles of market co-evolution.

The Lux and Lux and Marchesi Model

Lux [27] and Lux and Marchesi [28] propose a model to endogenously explain the heavy tail distribution of returns and the clustering of volatility. Both of these phenomena emerge in the Lux model as soon as one assumes that in addition to the fundamentalists there are also chartists in the model. Lux and Marchesi [28] further divide the chartists into optimists (buyers) and pessimists (sellers). The market fluctuations are driven and amplified by the fluctuations in the various populations: chartists converting into fundamentalists, pessimists into optimists, etc.

In the Lux and Marchesi model the stock's fundamental value is exogenously determined. The fluctuations of the fundamental value are inputted exogenously as a white noise process in the logarithm of the value. The market price is determined by investors' demands and by the market clearance condition.

Lux and Marchesi consider three types of traders:

- **Fundamentalists** observe the fundamental value of the stock. They anticipate that the price will eventually converge to the fundamental value, and their demand for shares is proportional to the difference between the market price and the fundamental value.
- **Chartists** look more at the present trends in the market price rather than at fundamental economic values; the chartists are divided into
- **Optimists** (they buy a fixed amount of shares per unit time)
- **Pessimists** (they sell shares).

Transitions between these three groups (optimists, pessimists, fundamentalists) happen with probabilities depending on the market dynamics and on the present numbers of traders in each of the three classes:

- **The transition probabilities of chartists** depend on the majority opinion (through an "opinion index" measuring the relative number of optimists minus the relative number of pessimists) and on the actual price trend (the current time derivative of the current market price), which determines the relative profit of the various strategies.

- **The fundamentalists decide to turn into chartists** if the profits of the later become significantly larger than their own, and vice versa (the detailed formulae used by Lux and Marchesi are inspired from the exponential transition probabilities governing statistical mechanics physical systems).

  The main results of the model are:

  - No long-term deviations between the current market price and the fundamental price are observed.
  - The deviations from the fundamental price, which do occur, are unsystematic.
  - In spite of the fact that the variations of the fundamental price are normally distributed, the variations of the market price (the market returns) are not. In particular the returns exhibit a frequency of extreme events which is higher than expected for a normal distribution. The authors emphasize the amplification role of the market that transforms the input normal distribution of the fundamental value variations into a leptokurtotic (heavy tailed) distribution of price variation, which is encountered in the actual financial data.
  - clustering of volatility.

The authors explain the volatility clustering (and as a consequence, the leptokurticity) by the following mechanism. In periods of high volatility, the fundamental information is not very useful to insure profits, and a large fraction of the agents become chartists. The opposite is true in quiet periods when the actual price is very close to the fundamental value. The two regimes are separated by a threshold in the number of chartist agents. Once this threshold is approached (from below) large fluctuations take place which further increase the number of chartists. This destabilization is eventually dampened by the energetic intervention of the fundamentalists when the price deviates too much from the fundamental value. The authors compare this temporal instability with the on-off intermittence encountered in certain physical systems. According to Egenter et al. [5], the fraction of chartists in the Lux Marchesi model goes to zero as the total number of traders goes to infinity, when the rest of the parameters are kept constant.

Illustration with the LLS Model

The purpose of this section is to give a more detailed "hands on" example of the agent based approach, and to discuss some of the practical dilemmas arising when implementing this approach, by focusing on one specific model. We will focus on the so called LLS Model of the stock market (for more detail, and various versions of the model, see [11,17,22,23,24,25]. This section is based on the presentation of the LLS Model in Chap. 7 of [24]).

Background

Real life investors differ in their investment behavior from the investment behavior of the idealized representative rational investor assumed in most economic and financial models. Investors differ one from the other in their preferences, their investment horizon, the information at their disposal, and their interpretation of this information. No financial economist seriously doubts these observations. However, modeling the empirically and experimentally documented investor behavior and the heterogeneity of investors is very difficult and in most cases practically impossible to do within an analytic framework. For instance, the empirical and experimental evidence suggests that most investors are characterized by Constant Relative Risk Aversion (CRRA), which implies a power (myopic) utility function (see Eq. (2) below). However, for a general distribution of returns it is impossible to obtain an analytic solution for the portfolio optimization problem of investors with these preferences. Extrapolation of future returns from past returns, biased probability weighting, and partial deviations from rationality are also all experimentally documented but difficult to incorporate in an analytical setting. One is then usually forced to make the assumptions of rationality and homogeneity (at least in some dimension) and to make unrealistic assumptions regarding investors' preferences, in order to obtain a model with a tractable solution. The hope in these circumstances is that the model will capture the essence of the system under investigation, and will serve as a useful benchmark, even though some of the underlying assumptions are admittedly false.

Most homogeneous rational agent models lead to the following predictions: no trading volume, zero autocorrelation of returns, and price volatility which is equal to or lower than the volatility of the "fundamental value" of the stock (defined as the present value of all future dividends, see [37]). However, the empirical evidence is very different:

- Trading volume can be extremely heavy [1,14].
- Stock returns exhibit short-run momentum (positive autocorrelation) and long-run mean reversion (negative autocorrelation) [7,13,21,31].

- Stock returns are excessively volatile relative to the dividends [37].

As most standard rational-representative-agent models cannot explain these empirical findings, these phenomena are known as"anomalies" or "puzzles". Can these "anomalies" be due to elements of investors' behavior which are unmodeled in the standard rational-representative-agent models, such as the experimentally documented deviations of investors' behavior from rationality and/or the heterogeneity of investors? The agent based simulation approach offers us a tool to investigate this question. The strength of the agent based simulation approach is that since it is not restricted to the scope of analytical methods, one is able to investigate virtually any imaginable investor behavior and market structure. Thus, one can study models which incorporate the experimental findings regarding the behavior of investors, and evaluate the effects of various behavioral elements on market dynamics and asset pricing.

The LLS model incorporates some of the main empirical findings regarding investor behavior, and we employ this model in order to study the effect of each element of investor behavior on asset pricing and market dynamics. We start out with a benchmark model in which all of the investors are rational, informed and identical, and then, one by one, we add elements of heterogeneity and deviations from rationality to the model in order to study their effects on the market dynamics.

In the benchmark model all investors are Rational, Informed and Identical (RII investors). This is, in effect, a "representative agent" model. The RII investors are informed about the dividend process, and they rationally act to maximize their expected utility. The RII investors make investment decisions based on the present value of future cash flows. They are essentially fundamentalists who evaluate the stock's fundamental value and try to find bargains in the market. The benchmark model in which all investors are RII yields results which are typical of most rational-representative-agent models: in this model prices follow a random walk, there is no excess volatility of the prices relative to the volatility of the dividend process, and since all agents are identical, there is no trading volume.

After describing the properties of the benchmark model, we investigate the effects of introducing various elements of investor behavior which are found in laboratory experiments but are absent in most standard models. We do so by adding to the model a minority of investors who do not operate like the RII investors. These investors are Efficient Market Believers (EMB from now on). The EMBs are investors who believe that the price of the stock reflects all of the currently available information about the stock. As a consequence, they do not try to time the market or to buy bargain stocks. Rather, their investment decision is reduced to the optimal diversification problem. For this portfolio optimization, the *ex-ante* return distribution is required. However, since the *ex-ante* distribution is unknown, the EMB investors use the *ex-post* distribution in order to estimate the *ex-ante* distribution. It has been documented that in fact, many investors form their expectations regarding the future return distribution based on the distribution of past returns.

There are various ways to incorporate the investment decisions of the EMBs. This stems from the fact that there are different ways to estimate the *ex-ante* distribution from the *ex-post* distribution. How far back should one look at the historical returns? Should more emphasis be given to more recent returns? Should some "outlier" observations be filtered out? etc. Of course, there are no clear answers to these questions, and different investors may have different ways of forming their estimation of the *ex-ante* return distribution (even though they are looking at the same series of historical returns). Moreover, some investors may use the objective *ex-post* probabilities when constructing their estimation of the *ex-ante* distribution, whereas others may use biased subjective probability weights. In order to build the analysis step-by-step we start by analyzing the case in which the EMB population is homogeneous, and then introduce various forms of heterogeneity into this population.

An important issue in market modeling is that of the degree of investors' rationality. Most models in economics and finance assume that people are fully rational. This assumption usually manifests itself as the maximization of an expected utility function by the individual. However, numerous experimental studies have shown that people deviate from rational decision-making [41,42,43,44,45]. Some studies model deviations from the behavior of the rational agent by introducing a sub-group of liquidity, or "noise", traders. These are traders that buy and sell stocks for reasons that are not directly related to the future payoffs of the financial asset - their motivation to trade arises from outside of the market (for example, a "noise trader's" daughter unexpectedly announces her plans to marry, and the trader sells stocks because of this unexpected need for cash). The exogenous reasons for trading are assumed random, and thus lead to random or "noise" trading (see [10]). The LLS model takes a different approach to the modeling of noise trading. Rather than dividing investors into the extreme categories of "fully rational" and "noise traders", the LLS model assumes that most investors try to act as rationally as they can, but are influenced by a multitude of factors causing them to deviate to some extent from the behavior that would have been optimal from their point of view. Namely, all investors are characterized by a utility function and act to maximize their expected utility; however, some investors may deviate to some extent from the optimal choice which maximizes their expected utility. These deviations from the optimal choice may be due to irrationality, inefficiency, liquidity constraints, or a combination of all of the above.

In the framework of the LLS model we examine the effects of the EMBs' deviations from rationality and their heterogeneity, relative to the benchmark model in which investors are informed, rational and homogeneous. We find that the behavioral elements which are empirically documented, namely, extrapolation from past returns, deviation from rationality, and heterogeneity among investors, lead to all of the following empirically documented "puzzles":

- Excess volatility
- Short-term momentum
- Longer-term return mean-reversion
- Heavy trading volume
- Positive correlation between volume and contemporaneous absolute returns
- Positive correlation between volume and lagged absolute returns

The fact that all these anomalies or "puzzles", which are hard to explain with standard rational-representative-agent models, are generated naturally by a simple model which incorporates the experimental findings regarding investor behavior and the heterogeneity of investors, leads one to suspect that these behavioral elements and the diversity of investors are a crucial part of the workings of the market, and as such they cannot be "assumed away". As the experimentally documented bounded-rational behavior and heterogeneity are in many cases impossible to analyze analytically, agent based simulation presents a very promising tool for investigating market models incorporating these elements.

The LLS Model

The stock market consists of two investment alternatives: a stock (or index of stocks) and a bond. The bond is assumed to be a riskless asset, and the stock is a risky asset. The stock serves as a proxy for the market portfolio (e. g., the Standard & Poors 500 index). The extension from one risky asset to many risky assets is possible; however, one stock (the index) is sufficient for our present analysis because we restrict ourselves to global market phenomena and do not wish to deal with asset allocation across several risky assets. Investors are allowed to revise their portfolio at given time points, i. e. we discuss a discrete time model.

The bond is assumed to be a riskless investment yielding a constant return at the end of each time period. The bond is in infinite supply and investors can buy from it as much as they wish at a given rate of r_f. The stock is in finite supply. There are N outstanding shares of the stock. The return on the stock is composed of two elements:

a) Capital Gain: If an investor holds a stock, any rise (fall) in the price of the stock contributes to an increase (decrease) in the investor's wealth.

b) Dividends: The company earns income and distributes dividends at the end of each time period. We denote the dividend per share paid at time t by D_t. We assume that the dividend is a stochastic variable following a multiplicative random walk, i. e., $\tilde{D}_t = D_{t-1}(1 + \tilde{z})$, where $\tilde{z}$ is a random variable with some probability density function $f(z)$ in the range $[z_1, z_2]$. (In order to allow for a dividend cut as well as a dividend increase we typically choose: $z_1 < 0, z_2 > 0$).

The total return on the stock in period t, which we denote by R_t is given by:

$$\tilde{R}_t = \frac{\tilde{P}_t + \tilde{D}_t}{P_{t-1}} , \qquad (1)$$

where $\tilde{P}_t$ is the stock price at time t.

All investors in the model are characterized by a von Neuman-Morgenstern utility function. We assume that all investors have a power utility function of the form:

$$U(W) = \frac{W^{1-\alpha}}{1-\alpha} , \qquad (2)$$

where α is the risk aversion parameter. This form of utility function implies Constant Relative Risk Aversion (CRRA). We employ the power utility function (Eq. (2)) because the empirical evidence suggests that relative risk aversion is approximately constant (see, for example [8,9,18,20]), and the power utility function is the unique utility function which satisfies the CRRA condition. Another implication of CRRA is that the optimal investment choice is independent of the investment horizon [33,34]. In other words, regardless of investors' actual investment horizon, they choose their optimal portfolio as though they are investing for a single period. The myopia property of the power utility function simplifies our analysis, as it allows us to assume that investors maximize their one-period-ahead expected utility.

We model two different types of investors: Rational, Informed, Identical (RII) investors, and Efficient Market Believers (EMB). These two investor types are described below.

Rational Informed Identical (RII) Investors RII investors evaluate the "fundamental value" of the stock as the discounted stream of all future dividends, and thus can also be thought of as "fundamentalists". They believe that the stock price may deviate from the fundamental value in the short run, but if it does, it will eventually converge to the fundamental value. The RII investors act according to the assumption of asymptotic convergence: if the stock price is low relative to the fundamental value they buy in anticipation that the underpricing will be corrected, and vice versa. We make the simplifying assumption that the RII investors believe that the convergence of the price to the fundamental value will occur in the next period, however, our results hold for the more general case where the convergence is assumed to occur some T periods ahead, with $T > 1$.

In order to estimate next period's return distribution, the RII investors need to estimate the distribution of next period's price, $\tilde{P}_{t+1}$, and of next period's dividend, $\tilde{D}_{t+1}$. Since they know the dividend process, the RII investors know that $\tilde{D}_{t+1} = D_t(1+\tilde{z})$ where $\tilde{z}$ is distributed according to $f(z)$ in the range $[z_1, z_2]$. The RII investors employ Gordon's dividend stream model in order to calculate the fundamental value of the stock:

$$P^f_{t+1} = \frac{E_{t+1}[\tilde{D}_{t+2}]}{k-g}\,, \tag{3}$$

where the superscript f stands for the *fundamental* value, $E_{t+1}[\tilde{D}_{t+2}]$ is the dividend corresponding to time $t+2$ as expected at time $t+1$, k is the discount factor or the expected rate of return demanded by the market for the stock, and g is the *expected* growth rate of the dividend, i. e., $g = E(\tilde{z}) = \int_{z_1}^{z_2} f(z)z\mathrm{d}z$.

The RII investors believe that the stock price may temporarily deviate from the fundamental value; however, they also believe that the price will eventually converge to the fundamental value. For simplification we assume that the RII investors believe that the convergence to the fundamental value will take place next period. Thus, the RII investors estimate P_{t+1} as:

$$P_{t+1} = P^f_{t+1}\,.$$

The expectation at time $t+1$ of $\tilde{D}_{t+2}$ depends on the realized dividend observed at $t+1$:

$$E_{t+1}[\tilde{D}_{t+2}] = D_{t+1}(1+g)\,.$$

Thus, the RII investors believe that the price at $t+1$ will be given by:

$$P_{t+1} = P^f_{t+1} = \frac{D_{t+1}(1+g)}{k-g}\,.$$

At time t, D_t is known, but D_{t+1} is not; therefore P^f_{t+1} is also not known with certainty at time t. However, given D_t, the RII investors know the distribution of $\tilde{D}_{t+1}$:

$$\tilde{D}_{t+1} = D_t(1+\tilde{z}),$$

where $\tilde{z}$ is distributed according to the known $f(z)$. The realization of $\tilde{D}_{t+1}$ determines P^f_{t+1}. Thus, at time t, RII investors believe that P_{t+1} is a random variable given by:

$$\tilde{P}_{t+1} = \tilde{P}^f_{t+1} = \frac{D_t(1+\tilde{z})(1+g)}{k-g}\,.$$

Notice that the RII investors face uncertainty regarding next period's price. In our model we assume that the RII investors are certain about the dividend growth rate g, the discount factor k, and the fact that the price will converge to the fundamental value next period. In this framework the only source of uncertainty regarding next period's price stems from the uncertainty regarding next period's dividend realization. More generally, the RII investors' uncertainty can result from uncertainty regarding any one of the above factors, or a combination of several of these factors. Any mix of these uncertainties is possible to investigate in the agent based simulation framework, but very hard, if not impossible, to incorporate in an analytic framework. As a consequence of the uncertainty regarding next period's price and of their risk aversion, the RII investors do not buy an infinite number of shares even if they perceive the stock as underpriced. Rather, they estimate the stock's next period's return distribution, and find the optimal mix of the stock and the bond which maximizes their expected utility. The RII investors estimate next period's return on the stock as:

$$\tilde{R}_{t+1} = \frac{\tilde{P}_{t+1} + \tilde{D}_{t+1}}{P_t} = \frac{\frac{D_t(1+\tilde{z})(1+g)}{k-g} + D_t(1+\tilde{z})}{P_t}\,, \tag{4}$$

where $\tilde{z}$, the next year growth in the dividend, is the source of uncertainty. The demands of the RII investors for the stock depend on the price of the stock. For any *hypothetical* price P_h investors calculate the proportion of their wealth x they should invest in the stock in order to maximize their expected utility. The RII investor i believes that if she invests a proportion x of her wealth in the stock at time t, then at time $t+1$ her wealth will be:

$$\tilde{W}^i_{t+1} = W^i_h[(1-x)(1+r_f) + x\tilde{R}_{t+1}], \tag{5}$$

where $\tilde{R}_{t+1}$ is the return on the stock, as given by Eq. (1), and W^i_h is the wealth of investor i at time t given that the stock price at time t is P_h.

If the price in period t is the hypothetical price P_h, the $t+1$ expected utility of investor i is the following function of her investment proportion in the stock, x:

$$EU(\tilde{W}^i_{t+1}) = EU\left(W^i_h\left[(1-x)(1+r_f) + x\tilde{R}_{t+1}\right]\right). \quad (6)$$

Substituting $\tilde{R}_{t+1}$ from Eq. (4), using the power utility function (Eq. (2)), and substituting the hypothetical price P_h for P_t, the expected utility becomes the following function of x:

$$EU(\tilde{W}^i_{t+1}) = \frac{(W^i_h)^{1-\alpha}}{1-\alpha} \int_{z_1}^{z_2} \left[(1-x)(1+r_f) + x\left(\frac{\frac{D_t(1+z)(1+g)}{k-g} + D_t(1+z)}{P_h}\right)\right]^{1-\alpha} f(z)\mathrm{d}z\,, \quad (7)$$

where the integration is over all possible values of z. In the agent based simulation framework, this expression for the expected utility, and the optimal investment proportion x, can be solved numerically for any general choice of distribution $f(z)$. For the sake of simplicity we restrict the present analysis to the case where $\tilde{z}$ is distributed uniformly in the range $[z_1, z_2]$. This simplification leads to the following expression for the expected utility:

$$\begin{aligned} &EU(\tilde{W}^i_{t+1}) \\ &= \frac{(W^i_h)^{1-\alpha}}{(1-\alpha)(2-\alpha)} \frac{1}{(z_2-z_1)} \left(\frac{k-g}{k+1}\right) \frac{P_h}{xD_t} \\ &\left\{\left[(1-x)(1+r_f) + \frac{x}{P_h}\left(\frac{k+1}{k-g}\right) D_t(1+z_2)\right]^{(2-\alpha)}\right. \\ &\left. -\left[(1-x)(1+r_f) + \frac{x}{P_h}\left(\frac{k+1}{k-g}\right) D_t(1+z_1)\right]^{(2-\alpha)}\right\} \end{aligned} \quad (8)$$

For any hypothetical price P_h, each investor (numerically) finds the optimal proportion x_h which maximizes his/her expected utility given by Eq. (8). Notice that the optimal proportion, x_h, is independent of the wealth, W^i_h. Thus, if all RII investors have the same degree of risk aversion, α, they will have the same optimal investment proportion in the stock, regardless of their wealth. The number of shares demanded by investor i at the hypothetical price P_h is given by:

$$N^i_h(P_h) = \frac{x^i_h(P_h)W^i_h(P_h)}{P_h}. \quad (9)$$

Efficient Market Believers (EMB) The second type of investors in the LLS model are EMBs. The EMBs believe in market efficiency - they believe that the stock price accurately reflects the stock's fundamental value. Thus, they do not try to time the market or to look for "bargain" stocks. Rather, their investment decision is reduced to the optimal diversification between the stock and the bond. This diversification decision requires the *ex-ante* return distribution for the stock, but as the *ex-ante* distribution is not available, the EMBs assume that the process generating the returns is fairly stable, and they employ the *ex-post* distribution of stock returns in order to estimate the *ex-ante* return distribution.

Different EMB investors may disagree on the optimal number of *ex-post* return observations that should be employed in order to estimate the *ex-ante* return distribution. There is a trade-off between using more observations for better statistical inference, and using a smaller number of only more recent observations, which are probably more representative of the *ex-ante* distribution. As in reality, there is no "recipe" for the optimal number of observations to use. EMB investor i believes that the m^i most recent returns on the stock are the best estimate of the *ex-ante* distribution. Investors create an estimation of the *ex-ante* return distribution by assigning an equal probability to each of the m^i most recent return observations:

$$Prob^i(\tilde{R}_{t+1} = R_{t-j}) = \frac{1}{m^i} \quad \text{for } j = 1, \ldots, m^i \quad (10)$$

The expected utility of EMB investor i is given by:

$$\begin{aligned} &EU(W^i_{t+1}) \\ &= \frac{(W^i_h)^{1-\alpha}}{(1-\alpha)} \frac{1}{m^i} \sum_{j=1}^{m^i} \left[(1-x)(1+r_f) + xR_{t-j}\right]^{1-\alpha}, \end{aligned} \quad (11)$$

where the summation is over the set of m^i most recent *ex-post* returns, x is the proportion of wealth invested in the stock, and as before W^i_h is the wealth of investor i at time t given that the stock price at time t is P_h. Notice that W^i_h does not change the optimal diversification policy, i. e., x. Given a set of m^i past returns, the optimal portfolio for the EMB investor i is an investment of a proportion x^{*i} in the stock and (1- x^{*i}) in the bond, where x^{*i} is the proportion which maximizes the above expected utility (Eq. (11)) for investor i. Notice that x^{*i} generally cannot be solved for analytically. However, in the agent based simulation framework this does not constitute a problem, as one can find x^{*i} numerically.

Deviations from Rationality Investors who are efficient market believers, and are rational, choose the investment proportion x^* which maximizes their expected utility. However, many empirical studies have shown that the behavior of investors is driven not only by rational expected utility maximization but by a multitude of other factors (see, for example, [34,41,42,43,44]). Deviations from the optimal rational investment proportion can be due to the cost of resources which are required for the portfolio optimization: time, access to information, computational power, etc., or due to exogenous events (for example, an investor plans to revise his portfolio, but gets distracted because his car breaks down). We assume that the different factors causing the investor to deviate from the optimal investment proportion x^* are random and uncorrelated with each other. By the central limit theorem, the aggregate effect of a large number of random uncorrelated influences is a normally distributed random influence, or "noise". Hence, we model the effect of all the factors causing the investor to deviate from his optimal portfolio by adding a normally distributed random variable to the optimal investment proportion. To be more specific, we assume:

$$x^i = x^{*i} + \tilde{\varepsilon}^i , \tag{12}$$

where $\tilde{\varepsilon}^i$ is a random variable drawn from a truncated normal distribution with mean zero and standard deviation σ. Notice that noise is investor-specific, thus, $\tilde{\varepsilon}^i$ is drawn separately and independently for each investor.

The noise can be added to the decision-making of the RII investors, the EMB investors, or to both. The results are not much different with these various approaches. Since the RII investors are taken as the benchmark of rationality, in this chapter we add the noise only to the decision-making of the EMB investors.

Market Clearance The number of shares demanded by each investor is a monotonically decreasing function of the hypothetical price P_h (see [24]). As the total number of outstanding shares is N, the price of the stock at time t is given by the market clearance condition: P_t is the unique price at which the total demand for shares is equal to the total supply, N:

$$\sum_i N_h^i(P_t) = \sum_i \frac{x_h(P_t) W_h^i(P_t)}{P_t} = N \,, \tag{13}$$

where the summation is over all the investors in the market, RII investors as well as EMB investors.

Agent Based Simulation The market dynamics begin with a set of initial conditions which consist of an initial stock price P_0, an initial dividend D_0, the wealth and number of shares held by each investor at time $t = 0$, and an initial "history" of stock returns. As will become evident, the general results do not depend on the initial conditions. At the first period ($t = 1$), interest is paid on the bond, and the time 1 dividend $\tilde{D}_1 = D_0(1 + \tilde{z})$ is realized and paid out. Then investors submit their demand orders, $N_h^i(P_h)$, and the market clearing price P_1 is determined. After the clearing price is set, the new wealth and number of shares held by each investor are calculated. This completes one time period. This process is repeated over and over, as the market dynamics develop.

We would like to stress that even the simplified benchmark model, with only RII investors, is impossible to solve analytically. The reason for this is that the optimal investment proportion, $x_h(P_h)$, cannot be calculated analytically. This problem is very general and it is encountered with almost any choice of utility function and distribution of returns. One important exception is the case of a negative exponential utility function and normally distributed returns. Indeed, many models make these two assumptions for the sake of tractability. The problem with the assumption of negative exponential utility is that it implies Constant Absolute Risk Aversion (CARA), which is very unrealistic, as it implies that investors choose to invest the same dollar amount in a risky prospect *independent of their wealth*. This is not only in sharp contradiction to the empirical evidence, but also excludes the investigation of the two-way interaction between wealth and price dynamics, which is crucial to the understanding of the market.

Thus, one contribution of the agent based simulation approach is that it allows investigation of models with realistic assumptions regarding investors' preferences. However, the main contribution of this method is that it permits us to investigate models which are much more complex (and realistic) than the benchmark model, in which all investors are RII. With the agent based simulation approach one can study models incorporating the empirically and experimentally documented investors' behavior, and the heterogeneity of investors.

Results of the LLS Model

We begin by describing the benchmark case where all investors are rational and identical. Then we introduce to the market EMB investors and investigate their affects on the market dynamics.

Benchmark Case: Fully Rational and Identical Agents In this benchmark model all investors are RII: rational, informed and identical. Thus, it is not surprising that the benchmark model generates market dynamics which are typical of homogeneous rational agent models:

No Volume All investors in the model are identical; they therefore always agree on the optimal proportion to invest in the stock. As a consequence, all the investors always achieve the same return on their portfolio. This means that at any time period the ratio between the wealth of any two investors is equal to the ratio of their initial wealths, i. e.:

$$\frac{W_t^i}{W_t^j} = \frac{W_0^i}{W_0^j} . \tag{14}$$

As the wealth of investors is always in the same proportion, and as they always invest the same fraction of their wealth in the stock, the number of shares held by different investors is also always in the same proportion:

$$\frac{N_t^i}{N_t^j} = \frac{\frac{x_t W_t^i}{P_t}}{\frac{x_t W_t^j}{P_t}} = \frac{W_t^i}{W_t^j} = \frac{W_0^i}{W_0^j} . \tag{15}$$

Since the total supply of shares is constant, this implies that each investor always holds the same number of shares, and there is *no trading volume* (the number of shares held may vary from one investor to the other as a consequence of different initial endowments).

Log-Prices Follow a Random Walk In the benchmark model all investors believe that next period's price will converge to the fundamental value given by the discounted dividend model (Eq. (3)). Therefore, the actual stock price is always close to the fundamental value. The fluctuations in the stock price are driven by fluctuations in the fundamental value, which in turn are driven by the fluctuating dividend realizations. As the dividend fluctuations are (by assumption) uncorrelated over time, one would expect that the price fluctuations will also be uncorrelated. To verify this intuitive result, we examine the return autocorrelations in simulations of the benchmark model.

Let us turn to the simulation of the model. We first describe the parameters and initial conditions used in the simulation, and then report the results. We simulate the benchmark model with the following parameters:

- Number of investors = 1000
- Risk aversion parameter α= 1.5. This value roughly conforms with the estimate of the risk aversion parameter found empirically and experimentally.
- Number of shares = 10,000.
- We take the time period to be a quarter, and accordingly we choose:
- Riskless interest rate $r_f = 0.01$.
- Required rate of return on stock $k = 0.04$.
- Maximal one-period dividend decrease z_1 = -0.07.
- Maximal one-period dividend growth z_2 = 0.10.
- $\tilde{z}$ is uniformly distributed between these values. Thus, the average dividend growth rate is $g = (z_1 + z_2)/2 = 0.015$.

Initial Conditions: Each investor is endowed at time $t = 0$ with a total wealth of $1000, which is composed of 10 shares worth an initial price of $50 per share, and $500 in cash. The initial quarterly dividend is set at $0.5 (for an annual dividend yield of about 4%). As will soon become evident, the dynamics are not sensitive to the particular choice of initial conditions.

Figure 1 shows the price dynamics in a typical simulation with these parameters (simulations with the same parameters differ one from the other because of the different random dividend realizations). Notice that the vertical axis in this figure is logarithmic. Thus, the roughly constant slope implies an approximately exponential price growth, or an approximately constant *average* return.

The prices in this simulation seem to fluctuate randomly around the trend. However, Fig. 1 shows only one simulation. In order to have a more rigorous analysis we perform many independent simulations, and employ statistical tools. Namely, for each simulation we calculate the autocorrelation of returns. We perform a univariate regression of the return in time t on the return on time $t-j$:

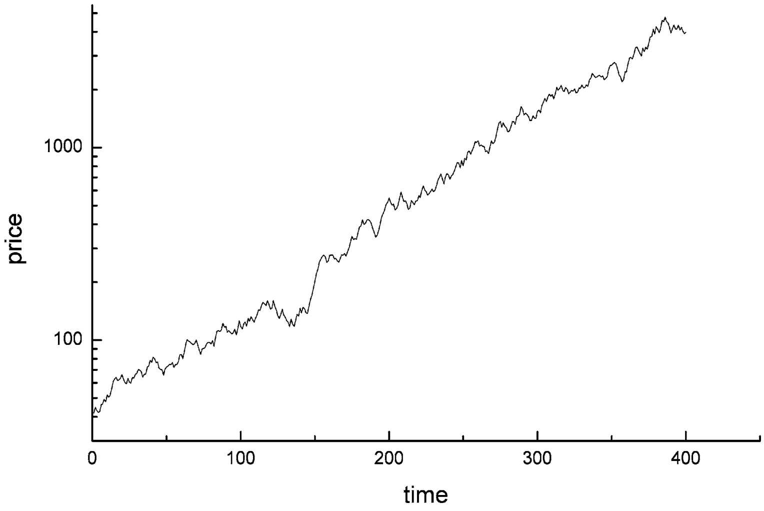
$$R_t = \alpha_j + \beta_j R_{t-j} + \varepsilon ,$$

where R_t is the return in period t, and j is the lag. The autocorrelation of returns for lag j is defined as:

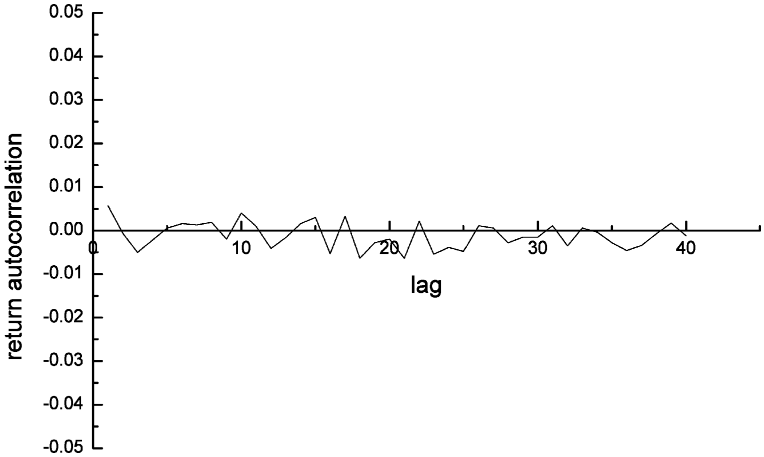
$$\rho_j = \frac{cov(R_t, R_{t-j})}{\hat{\sigma}^2(R)} ,$$

and it is estimated by $\hat{\beta}$. We calculate the autocorrelation for different lags, $j = 1, \ldots 40$. Figure 2 shows the average autocorrelation as a function of the lag, calculated over 100 independent simulations. It is evident both from the figure that the returns are uncorrelated in the benchmark model, conforming with the random-walk hypothesis.

No Excess Volatility Since the RII investors believe that the stock price will converge to the fundamental value next period, in the benchmark model prices are always close to the fundamental value given by the discounted dividend stream. Thus, we do not expect prices to be more volatile than the value of the discounted dividend stream. For a formal test of excess volatility we follow the technique in [37]. For each time period we calculate the actual



Agent Based Computational Economics, Figure 1
Price Dynamics in the Benchmark Model



Agent Based Computational Economics, Figure 2
Return Autocorrelation in Benchmark Model

price P_t, and the fundamental value of discounted dividend stream, P_t^f, as in Eq. (3). Since prices follow an upward trend, in order to have a meaningful measure of the volatility, we must detrend these price series. Following Shiller, we run the regression:

$$\ln P_t = bt + c + \varepsilon_t \,, \tag{16}$$

in order to find the average exponential price growth rate (where b and c are constants). Then, we define the detrended price as: $p_t = P_t / e^{\hat{b}t}$. Similarly, we define the detrended value of the discounted dividend stream p_t^f, and compare $\sigma(p_t)$ with $\sigma(p_t^f)$. For 100 1000-period simulations we find an average $\sigma(p_t)$ of 22.4, and an average $\sigma(p_t^f)$ of 22.9. As expected, the actual price and the fundamental value have almost the same volatility.

To summarize the results obtained for the benchmark model, we find that when all investors are assumed to be rational, informed and identical, we obtain results which are typical of rational-representative-agent models: no volume, no return autocorrelations, and no excess volatility. We next turn to examine the effect of introducing into the market EMB investors, which model empirically and experimentally documented elements of investors' behavior.

The Introduction of a Small Minority of EMB Investors
In this section we will show that the introduction of a small minority of heterogeneous EMB investors generates many

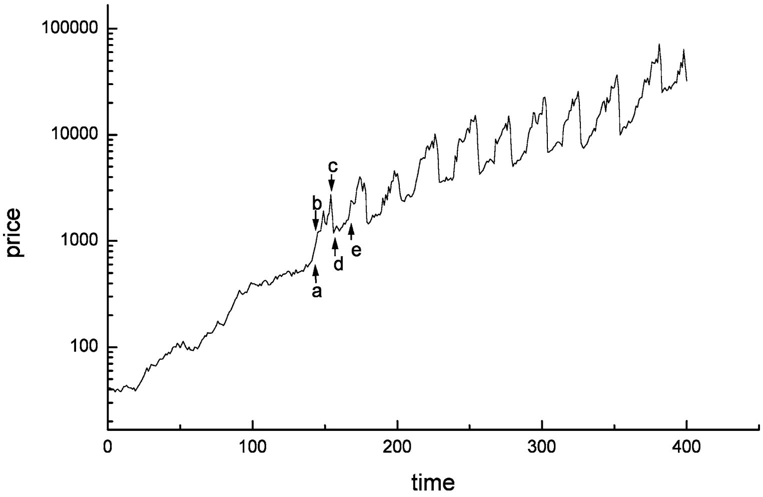
of the empirically observed market "anomalies" which are absent in the benchmark model, and indeed, in most other rational-representative-agent models. We take this as strong evidence that the "non-rational" elements of investor behavior which are documented in experimental studies, and the heterogeneity of investors, both of which are incorporated in the LLS model, are crucial to understanding the dynamics of the market.

In presenting the results of the LLS model with EMB investors we take an incremental approach. We begin by describing the results of a model with a small sub-population of *homogeneous* EMB believers. This model produces the above mentioned market "anomalies"; however, it produces unrealistic cyclic market dynamics. Thus, this model is presented both for analyzing the source of the "anomalies" in a simplified setting, and as a reference point with which to compare the dynamics of the model with a *heterogeneous* EMB believer population.

We investigate the effects of investors' heterogeneity by first analyzing the case in which there are two types of EMBs. The two types differ in the method they use to estimate the *ex-ante* return distribution. Namely, the first type looks at the set of the last m_1 *ex-post* returns, whereas the second type looks at the set of the last m_2 *ex-post* returns. It turns out that the dynamics in this case are much more complicated than a simple "average" between the case where all EMB investors have m_1 and the case where all EMB investors have m_2. Rather, there is a complex non-linear interaction between the two EMB sub-populations. This implies that the heterogeneity of investors is a very important element determining the market dynamics, an element which is completely absent in representative-agent models.

Finally, we present the case where there is an entire spectrum of EMB investors differing in the number of *ex-post* observations they take into account when estimating the *ex-ante* distribution. This general case generates very realistic-looking market dynamics with all of the above mentioned market anomalies.

Homogeneous Sub-Population of EMBs When a very small sub-population of EMB investors is introduced to the benchmark LLS model, the market dynamics change dramatically. Figure 3 depicts a typical price path in a simulation of a market with 95% RII investors and 5% EMB investors. The EMB investors have $m = 10$ (i. e., they estimate the *ex-ante* return distribution by observing the set of the last 10 *ex-post* returns). σ, the standard deviation of the random noise affecting the EMBs' decision making is taken as 0.2. All investors, RII and EMB alike, have the same risk aversion parameter $\alpha = 1.5$ (as before). In the first 150 trading periods the price dynamics look very similar to the typical dynamics of the benchmark model. However, after the first 150 or so periods the price dynamics change. From this point onwards the market is characterized by periodic booms and crashes. Of course, Fig. 3 describes only one simulation. However, as will become evident shortly, different simulations with the same parameters may differ in detail, but the pattern is general: at some stage (not necessarily after 150 periods) the EMB investors



Agent Based Computational Economics, Figure 3
5% of Investors are Efficient Market Believers, 95% Rational Informed Investors

induce cyclic price behavior. It is quite astonishing that such a small minority of only 5% of the investors can have such a dramatic impact on the market.

In order to understand the periodic booms and crashes let us focus on the behavior of the EMB investors. After every trade, the EMB investors revise their estimation of the *ex-ante* return distribution, because the set of *ex-post* returns they employ to estimate the *ex-ante* distribution changes. Namely, investors add the latest return generated by the stock to this set and delete the oldest return from this set. As a result of this update in the estimation of the *ex-ante* distribution, the optimal investment proportion x^* changes, and EMB investors revise their portfolios at next period's trade. During the first 150 or so periods, the informed investors control the dynamics and the returns fluctuate randomly (as in the benchmark model). As a consequence, the investment proportion of the EMB investors also fluctuates irregularly. Thus, during the first 150 periods the EMB investors do not effect the dynamics much. However, at point **a** the dynamics change qualitatively (see Fig. 3). At this point, a relatively high dividend is realized, and as a consequence, a relatively high return is generated. This high return leads the EMB investors to increase their investment proportion in the stock at the next trading period. This increased demand of the EMB investors is large enough to effect next period's price, and thus a second high return is generated. Now the EMB investors look at a set of *ex-post* returns with two high returns, and they increase their investment proportion even further. Thus, a positive feedback loop is created.

Notice that as the price goes up, the informed investors realize that the stock is overvalued relative to the fundamental value P^f and they decrease their holdings in the stock. However, this effect does not stop the price increase and break the feedback loop because the EMB investors continue to buy shares aggressively. The positive feedback loop pushes the stock price further and further up to point **b**, at which the EMBs are invested 100% in the stock. At point **b** the positive feedback loop "runs out of gas". However, the stock price remains at the high level because the EMB investors remain fully invested in the stock (the set of past m=10 returns includes at this stage the very high returns generated during the "boom" – segment **a**–**b** in Fig. 3).

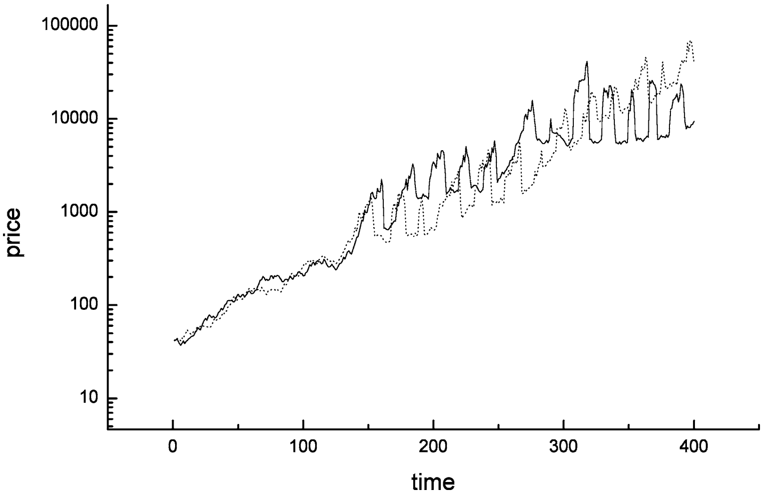
When the price is at the high level (segment **b–c**), the dividend yield is low, and as a consequence, the returns are generally low. As time goes by and we move from point **b** towards point **c**, the set of $m = 10$ last returns gets filled with low returns. Despite this fact, the extremely high returns generated in the boom are also still in this set, and they are high enough to keep the EMB investors fully invested. However, 10 periods after the boom, these extremely high returns are pushed out of the set of relevant *ex-post* returns. When this occurs, at point **c**, the EMB investors face a set of low returns, and they cut their investment proportion in the stock sharply. This causes a dramatic crash (segment **c–d**). Once the stock price goes back down to the "fundamental" value, the informed investors come back into the picture. They buy back the stock and stop the crash.

The EMB investors stay away from the stock as long as the *ex-post* return set includes the terrible return of the crash. At this stage the informed investors regain control of the dynamics and the stock price remains close to its fundamental value. 10 periods after the crash the extremely negative return of the crash is excluded from the *ex-post* return set, and the EMB investors start increasing their investment proportion in the stock (point **e**). This drives the stock price up, and a new boom-crash cycle is initiated. This cycle repeats itself over and over almost periodically.

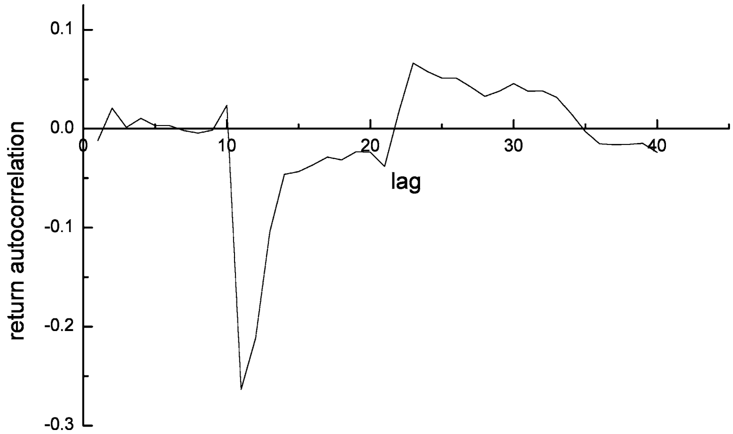
Figure 3 depicts the price dynamics of a single simulation. One may therefore wonder how general the results discussed above are. Figure 4 shows two more simulations with the same parameters but different dividend realizations. It is evident from this figure that although the simulations vary in detail (because of the different dividend realizations), the overall price pattern with periodic boom-crash cycles is robust.

Although these dynamics are very unrealistic in terms of the periodicity, and therefore the predictability of the price, they do shed light on the mechanism generating many of the empirically observed market phenomena. In the next section, when we relax the assumption that the EMB population is homogeneous with respect to m, the price is no longer cyclic or predictable, yet the mechanisms generating the market phenomena are the same as in this homogeneous EMB population case. The homogeneous EMB population case generates the following market phenomena:

Heavy Trading Volume As explained above, shares change hands continuously between the RII investors and the EMB investors. When a "boom" starts the RII investors observe higher ex-post returns and become more optimistic, while the RII investor view the stock as becoming overpriced and become more pessimistic. Thus, at this stage the EMBs buy most of the shares from the RIIs. When the stock crashes, the opposite is true: the EMBs are very pessimistic, but the RII investors buy the stock once it falls back to the fundamental value. Thus, there is substantial trading volume in this market. The average trading



Agent Based Computational Economics, Figure 4
Two More Simulations – same Parameters as Fig. 3, Different Divident Realizations



Agent Based Computational Economics, Figure 5
Return Autocorrelation 5%, Efficient Market Believers, $m = 10$

volume in a typical simulation is about 1000 shares per period, which are 10% of the total outstanding shares.

Autocorrelation of Returns The cyclic behavior of the price yields a very definite return autocorrelation pattern. The autocorrelation pattern is depicted graphically in Fig. 5. The autocorrelation pattern is directly linked to the length of the price cycle, which in turn are determined by m. Since the moving window of *ex-post* returns used to estimate the *ex-ante* distribution is $m = 10$ periods long, the price cycles are typically a little longer than 20 periods long: a cycle consists of the positive feedback loop (segment **a–b** in Fig. 3) which is about 2–3 periods long, the upper plateau (segment **b–c** in Fig. 3) which is about 10 periods long, the crash that occurs during one or two periods, and the lower plateau (segment **d–e** in Fig. 3) which is again about 10 periods long, for a total of about 23–25 periods. Thus, we expect positive autocorrelation for lags of about 23–25 periods, because this is the lag between one point and the corresponding point in the next (or previous) cycle. We also expect negative autocorrelation for lags of about 10–12 periods, because this is the lag between a boom and the following (or previous) crash, and vice versa. This is precisely the pattern we observe in Fig. 5.

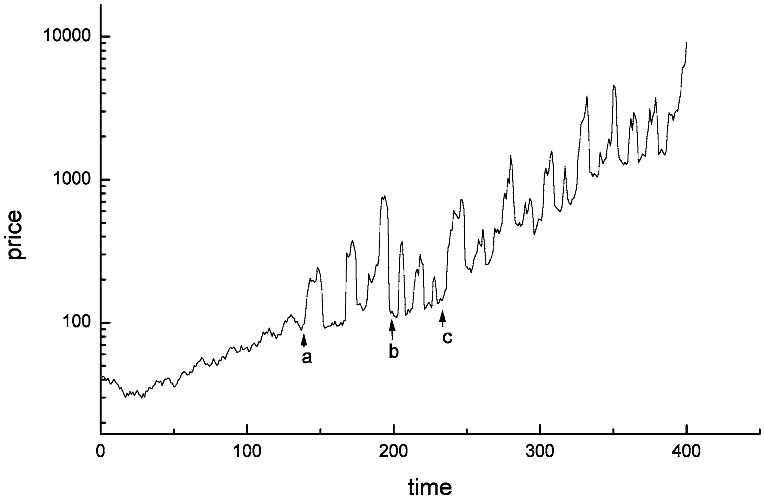
Excess Volatility The EMB investors induce large deviations of the price from the fundamental value. Thus, price fluctuations are caused not only by dividend fluctuations (as the standard theory suggests) but also by the endogenous market dynamics driven by the EMB investors. This "extra" source of fluctuations causes the price to be more volatile than the fundamental value P^f.

Indeed, for 100 1000-period independent simulations with 5% EMB investors we find an average $\sigma(p_t)$ of 46.4, and an average $\sigma(p_t^f)$ of 30.6; i. e., we have excess volatility of about 50%.

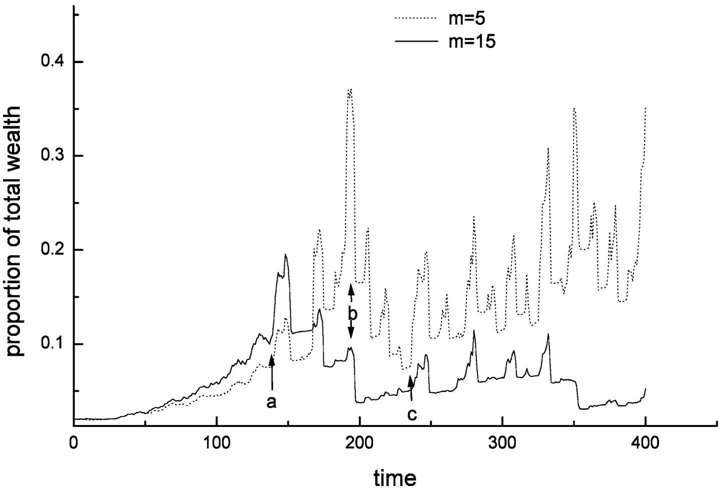
As a first step in analyzing the effects of heterogeneity of the EMB population, in the next section we examine the case of two types of EMB investors. We later analyze a model in which there is a full spectrum of EMB investors.

Two Types of EMBs One justification for using a representative agent in economic modeling is that although investors are heterogeneous in reality, one can model their collective behavior with one representative or "average" investor. In this section we show that this is generally not true. Many aspects of the dynamics result from the non-linear interaction between different investor types. To illustrate this point, in this section we analyze a very simple case in which there are only two types of EMB investors: one with $m = 5$ and the other with $m = 15$. Each of these two types consists of 2% of the investor population, and the remaining 96% are informed investors. The representative agent logic may tempt us to think that the resulting market dynamics would be similar to that of one "average" investor, i. e. an investor with $m = 10$. Figure 6 shows that this is clearly not the case. Rather than seeing periodic cycles of about 23–25 periods (which correspond to the average m of 10, as in Fig. 3), we see an irregular pattern. As before, the dynamics are first dictated by the informed investors. Then, at point **a**, the EMB investors with $m = 15$ induce cycles which are about 30 periods long. At point **b** there is a transition to shorter cycles induced by the $m = 5$ population, and at point **c** there is another transition back to longer cycles. What is going on?

These complex dynamics result from the non-linear interaction between the different sub-populations. The transitions from one price pattern to another can be partly understood by looking at the wealth of each sub-population. Figure 7 shows the proportion of the total wealth held by each of the two EMB populations (the remaining proportion is held by the informed investors). As seen in Fig. 7, the cycles which start at point **a** are dictated by the $m = 15$ rather than the $m = 5$ population, because at this stage the $m = 15$ population controls more of the wealth than the $m = 5$ population. However, after 3 cycles (at point **b**) the picture is reversed. At this point the $m = 5$ population is more powerful than the $m = 15$ population, and there is a transition to shorter boom-crash cycles. At point **c** the wealth of the two sub-populations is again almost equal, and there is another transition to longer cycles. Thus, the complex price dynamics can be partly understood from the wealth dynamics. But how are the wealth dynamics determined? Why does the $m = 5$ population become wealthier at point **b**, and why does it lose most of this advantage at point **c**? It is obvious that the wealth dynamics are influenced by the price dynamics, thus there is a complicated two-way interaction between the two. Although this interaction is generally very complex, some principle ideas about the mutual influence between the wealth and price patterns can be formulated. For example, a population that becomes



Agent Based Computational Economics, Figure 6
2% EMB $m = 5$, 2% EMB $m = 15$, 96% RII



Agent Based Computational Economics, Figure 7
Proportion of the total wealth held by the two EMB populations

dominant and dictates the price dynamics, typically starts under-performing, because it affects the price with its actions. This means pushing the price up when buying, and therefore buying high, and pushing the price down when selling. However, a more detailed analysis must consider the specific investment strategy employed by each population. For a more comprehensive analysis of the interaction between heterogeneous EMB populations see [25].

The two EMB population model generates the same market phenomena as did the homogeneous population case: heavy trading volume, return autocorrelations, and excess volatility. Although the price pattern is much less regular in the two-EMB-population case, there still seems to be a great deal of predictability about the prices. Moreover, the booms and crashes generated by this model are unrealistically dramatic and frequent. In the next section we analyze a model with a continuous spectrum of EMB investors. We show that this fuller heterogeneity of investors leads to very realistic price and volume patterns.

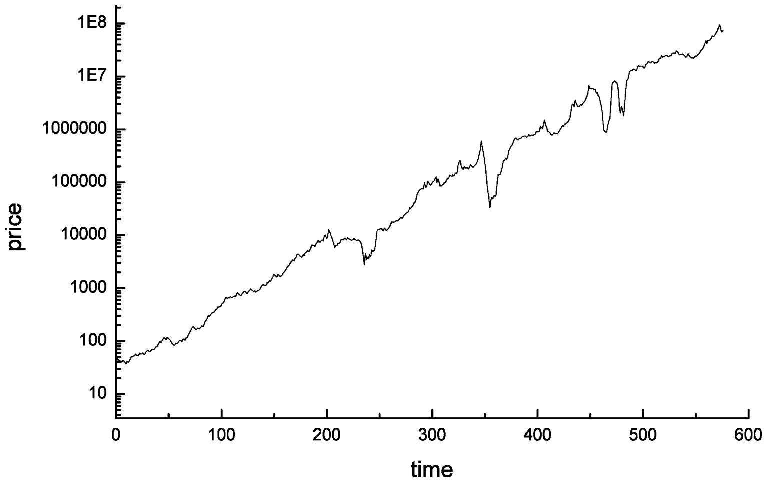
Full Spectrum of EMB Investors Up to this point we have analyzed markets with at most three different sub-populations (one RII population and two EMB populations). The market dynamics we found displayed the empirically observed market anomalies, but they were unrealistic in the magnitude, frequency, and semi-predictability of booms and crashes. In reality, we would expect not only two or three investor types, but rather an entire spectrum of investors. In this section we consider a model with a full spectrum of different EMB investors. It turns out that *more is different*. When there is an entire range of investors, the price dynamics become realistic: booms and crashes are not periodic or predictable, and they are also less frequent and dramatic. At the same time, we still obtain all of the market anomalies described before.

In this model each investor has a different number of *ex-post* observations which he utilizes to estimate the *ex-ante* distribution. Namely, investor i looks at the set of the m^i most recent returns on the stock, and we assume that m^i is distributed in the population according to a truncated normal distribution with average $\bar{m}$ and standard deviation σ_m (as $m \leq 0$ is meaningless, the distribution is truncated at $m = 0$).

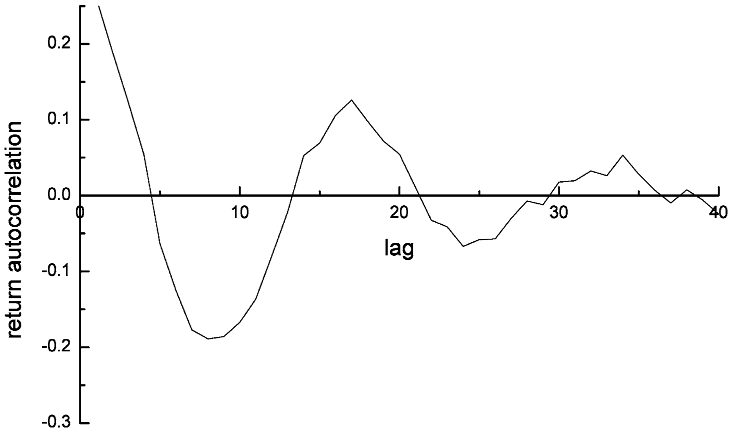
Figure 8 shows the price pattern of a typical simulation of this model. In this simulation 90% of the investors are RII, and the remaining 10% are heterogeneous EMB investors with $\bar{m} = 40$, and $\sigma_m = 10$. The price pattern seems very realistic with "smoother" and more irregular cycles. Crashes are dramatic, but infrequent and unpredictable.

The heterogeneous EMB population model generates the following empirically observed market phenomena:

Return Autocorrelation: Momentum and Mean-Reversion In the heterogeneous EMB population model trends are generated by the same positive feedback mechanism that generated cycles in the homogeneous case: high (low) returns tend to make the EMB investors more (less) aggressive, this generates more high (low) returns, etc. The difference between the two cases is that in the heterogeneous case there is a very complicated interaction between all the different investor sub-populations and as a result there are no distinct regular cycles, but rather, smoother and more irregular trends. There is no single cycle length –



Agent Based Computational Economics, Figure 8
Spectrum of Heterogeneous EMB Investors (10% EMB Investors, 90% RII Investors)



Agent Based Computational Economics, Figure 9
Return Autocorrelation – Heterogeneous EMB Population

the dynamics are a combination of many different cycles. This makes the autocorrelation pattern also smoother and more continuous. The return autocorrelations in the heterogeneous model are shown in Fig. 9. This autocorrelation pattern conforms with the empirical findings. In the short-run (lags 1–4) the autocorrelation is positive – this is the empirically documented phenomena known as momentum: in the short-run, high returns tend to be followed by more high returns, and low returns tend to be followed by more low returns. In the longer-run (lags 5–13) the autocorrelation is negative, which is known as mean-reversion. For even longer lags the autocorrelation eventually tends to zero. The short-run momentum, longer-run mean-reversion, and eventual diminishing autocorrelation creates the general "U-shape" which is found in empirical studies [7,13,31] and which is seen in Fig. 9.

Excess Volatility The price level is generally determined by the fundamental value of the stock. However, as in the homogeneous EMB population case, the EMB investors occasionally induce temporary departures of the price away from the fundamental value. These temporary departures from the fundamental value make the price more volatile than the fundamental value. Following Shiller's methodology we define the detrended price, p, and fundamental value, p^f. Averaging over 100 independent simulations we find $\sigma(p) = 27.1$ and $\sigma(p^f) = 19.2$, which is an excess volatility of 41% .

Heavy Volume As investors in our model have different information (the informed investors know the dividend process, while the EMB investors do not), and different ways of interpreting the information (EMB investors with different memory spans have different estimations regarding the *ex-ante* return distribution), there is a high level of trading volume in this model. The average trading volume in this model is about 1700 shares per period (17% of the total outstanding shares). As explained below, the volume is positively correlated with contemporaneous and lagged absolute returns.

Volume is Positively Correlated with Contemporaneous and Lagged Absolute Returns Investors revise their portfolios as a result of changes in their beliefs regarding the future return distribution. The changes in the beliefs can be due to a change in the current price, to a new dividend realization (in the case of the informed investors), or to a new observation of an *ex-post* return (in the case of the EMB investors). If all investors change their beliefs in the same direction (for example, if everybody becomes more optimistic), the stock price can change substantially with almost no volume – everybody would like to increase the proportion of the stock in his portfolio, this will push the price up, but a very small number of shares will change hands. This scenario would lead to zero or perhaps even negative correlation between the magnitude of the price change (or return) and the volume. However, the typical scenario in the LLS model is different. Typically, when a positive feedback trend is induced by the EMB investors, the opinions of the informed investors and the EMB investors change in opposite directions. The EMB investors see a trend of rising prices as a positive indication about the *ex-ante* return distribution, while the informed investors believe that the higher the price level is above the fundamental value, the more overpriced the stock is, and the harder it will eventually fall. The exact opposite holds for a trend of falling prices. Thus, price trends are typically interpreted differently by the two investor types, and therefore induce heavy trading volume. The more pronounced the trend, the more likely it is to lead to heavy volume, and at the same time, to large price changes which are due to the positive feedback trading on behalf of the EMB investors.

This explains not only the positive correlation between volume and contemporaneous absolute rates of return, but also the positive correlation between volume and lagged absolute rates of return. The reason is that the behavior of the EMB investors induces short-term positive return autocorrelation, or momentum (see above). That is, a large absolute return this period is associated not only with high volume this period, but also with a large absolute return next period, and therefore with high volume next period. In other words, when there is a substantial price increase (decrease), EMB investors become more (less) aggressive and the opposite happens to the informed traders. As we have seen before, when a positive feedback loop is started, the EMB investors are more dominant in determining the price, and therefore another large price increase (decrease) is expected next period. This large price change is likely to be associated with heavy trading volume as the opinions of the two populations diverge. Furthermore, this large increase (decrease) is expected to make the EMB investors even more optimistic (pessimistic) leading to another large price increase (decrease) and heavy volume next period.

In order to verify this relationship quantitatively, we regress volume on contemporaneous and lagged absolute rates of return for 100 independent simulations. We run the regressions:

$$\begin{aligned} V_t &= \alpha + \beta_C \,|R_t - 1| + \varepsilon_t\,, \quad \text{and} \\ V_t &= \alpha + \beta_L \,|R_{t-1} - 1| + \varepsilon_t\,, \end{aligned} \tag{17}$$

where V_t is the volume at time t and R_t is the total return on the stock at time t, and the subscripts C and L stand for contemporaneous and lagged. We find an average value of 870 for $\hat{\beta}_C$ with an average t-value of 5.0 and an average value of 886 for $\hat{\beta}_L$ with an average t-value of 5.1.

Discussion of the LLS Results The LLS model is an Agent Based Simulation model of the stock market which incorporates some of the fundamental experimental findings regarding the behavior of investors. The main nonstandard assumption of the model is that there is a small minority of investors in the market who are uninformed about the dividend process and who believe in market efficiency. The investment decision of these investors is reduced to the optimal diversification between the stock and the bond.

The LLS model generates many of the empirically documented market phenomena which are hard to explain in the analytical rational-representative-agent framework. These phenomena are:

- Short term momentum;
- Longer term mean reversion;
- Excess volatility;
- Heavy trading volume;
- Positive correlation between volume and contemporaneous absolute returns;
- Positive correlation between volume and lagged absolute returns;
- Endogenous market crashes.

The fact that so many "puzzles" are explained with a simple model built on a small number of empirically documented behavioral elements leads us to suspect that these behavioral elements are very important in understanding the workings of the market. This is especially true in light of the observations that a very small minority of the nonstandard bounded-rational investors can have a dramatic influence on the market, and that these investors are not wiped out by the majority of rational investors.

Summary and Future Directions

Standard economic models typically describe a world of homogeneous rational agents. This approach is the foundation of most of our present day knowledge in economic theory. With the Agent Based Simulation approach we can investigate a much more complex and "messy" world with different agent types, who employ different strategies to try to survive and prosper in a market with structural uncertainty. Agents can learn over time, from their own experience and from their observation about the performance of other agents. They co-evolve over time and as they do so, the market dynamics change continuously. This is a world view closer to biology, than it is to the "clean" realm of physical laws which classical economics has aspired to.

The Agent Based approach should not and can not replace the standard analytical economic approach. Rather, these two methodologies support and complement each other: When an analytical model is developed, it should become standard practice to examine the robustness of the model's results with agent based simulations. Similarly, when results emerge from agent based simulation, one should try to understand their origin and their generality, not only by running many simulations, but also by trying to capture the essence of the results in a simplified analytical setting (if possible).

Although the first steps in economic agent based simulations were made decades ago, economics has been slow and cautious to adopt this new methodology. Only in recent years has this field begun to bloom. It is my belief and hope that the agent based approach will prove as fruitful in economics as it has been in so many other branches of science.

Bibliography

Primary Literature

1. Admati A, Pfleiderer P (1988) A theory of intraday patterns: Volume and price variability. Rev Financ Stud 1:3–40
2. Arthur WB (1994) Inductive reasoning and bounded rationality (The El Farol problem). Am Econ Rev 84:406–411
3. Arthur WB, Holland JH, Lebaron B, Palmer RG, Tayler P (1997) Asset pricing under endogenous expectations in an artificial stock market. In: Arthur WB, Durlauf S, Lane D (eds) The economy as an evolving complex system II. Addison-Wesley, Redwood City
4. Brock WA, Hommes CA (1998) Heterogeneous beliefs and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
5. Egenter E, Lux T, Stauffer D (1999) Finite size effects in Monte Carlo Simulations of two stock market models. Physica A 268:250–256
6. Epstein JM, Axtell RL (1996) Complex adaptive systems. In: Growing artificial societies: Social science from the bottom up. MIT Press, Washington DC
7. Fama E, French K (1988) Permanent and temporary components of stock prices. J Political Econ 96:246–273
8. Friend I, Blume ME (1975) The demand for risky assets. Am Econ Rev 65:900–922
9. Gordon J, Paradis GE, Rorke CH (1972) Experimental evidence on alternative portfolio decision rules. Am Econ Rev 62(1):107–118
10. Grossman S, Stiglitz J (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70:393–408
11. Hellthaler T (1995) The influence of investor number on a microscopic market. Int J Mod Phys C 6:845–852
12. Hommes CH (2002) Modeling the stylized facts in finance through simple nonlinear adaptive systems. PNAS 99:7221–7228
13. Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. J Finance 48:65–91
14. Karpoff J (1987) The relationship between price changes and trading volume: A survey. J Financ Quantitative Anal 22:109–126
15. Kim GW, Markowitz HM (1989) Investment rules, margin, and market volatility. J Portf Manag 16:45–52
16. Kirman AP (1992) Whom or what does the representative agent represent? J Econ Perspectiv 6:117–136
17. Kohl R (1997) The influence of the number of different stocks on the Levy, Levy Solomon model. Int J Mod Phys C 8:1309–1316
18. Kroll Y, Levy H, Rapoport A (1988) Experimental tests of the separation theorem and the capital asset pricing model. Am Econ Rev 78:500–519
19. LeBaron B (2000) Agent-based computational finance: Suggested readings and early research. J Econ Dyn Control 24:679–702
20. Levy H (1994) Absolute and relative risk aversion: An experimental study. J Risk Uncertain 8:289–307
21. Levy H, Lim KC (1998) The economic significance of the cross-sectional autoregressive model: Further analysis. Rev Quant Finance Acc 11:37–51
22. Levy M, Levy H (1996) The danger of assuming homogeneous expectations. Financ Analyst J 52:65–70
23. Levy M, Levy H, Solomon S (1994) A microscopic model of the stock market: Cycles, booms, and crashes. Econs Lett 45:103–111
24. Levy M, Levy H, Solomon S (2000) Microscopic simulation of financial markets. Academic Press, San Diego
25. Levy M, Persky N, Solomon, S (1996) The complex dyn of a simple stock market model. Int J High Speed Comput 8:93–113

26. Lux T (1995) Herd behaviour, bubbles and crashes. Econ J 105:881
27. Lux T (1998) The socio-economic dynamics of speculative bubbles: Interacting agents, chaos, and the fat tails of returns distributions. J Econ Behav Organ 33:143–165
28. Lux T, Marchesi M (1999) Volatility clustering in financial markets: A micro-simulation of interacting agents. Nature 397: 498
29. Orcutt GH, Caldwell SB, Wertheimer R (1976) Policy exploration through microanalytic simulation. The Urban Institute, Washington DC
30. Palmer RG, Arthur WB, Holland JH, LeBaron B, Tayler P (1994) Artificial economic life: A simple model of a stock market. Physica D 75:264–274
31. Poterba JM, Summers LH (1988) Mean reversion in stock returns: Evidence and implications. J Financial Econs 22:27–59
32. Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Agent-based models of financial markets. Rep Prog Phys 70:409–450
33. Samuelson PA (1989) The judgement of economic science on rational portfolio management: Timing and long horizon effects. J Portfolio Manag 16:4–12
34. Samuelson PA (1994) The long term case for equities and how it can be oversold. J Portf Management 21:15–24
35. Sargent T (1993) Bounded rationality and macroeconomics. Oxford University Press, Oxford
36. Schelling TC (1978) Micro motives and macro behavior. Norton & Company, New York
37. Shiller RJ (1981) Do stock prices move too much to be justified by subsequent changes in dividends? Am Econ Rev 71: 421–436
38. Stauffer D, de Oliveira PMC, Bernardes AT (1999) Monte Carlo Simulation of volatility correlation in microscopic market model. Int J Theor Appl Finance 2:83–94
39. Tesfatsion L (2002) Agent-based computational economics: Growing economies from the bottom up. Artif Life 8:55–82
40. Tesfatsion L (2001) Special issue on agent-based computational economics. J Econ Dyn Control 25:281–293
41. Thaler R (ed) (1993) Advances in behavioral finance. Russel Sage Foundation, New York
42. Thaler R (1994) Quasi rational economics. Russel Sage Foundation, New York
43. Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–480
44. Tversky A, Kahneman D (1986) Rational choice and the framing of decision. J Bus 59(4):251–278
45. Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertain 5:297–323

Books and Reviews

Anderson PW, Arrow J, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Redwood City

Axelrod R (1997) The complexity of cooperation: Agent-based models of conflict and cooperation. Princeton University Press, Princeton

Moss de Oliveira S, de Oliveira H, Stauffer D (1999) Evolution, money, war and computers. BG Teubner, Stuttgart-Leipzig

Solomon S (1995) The microscopic representation of complex macroscopic phenomena. In: Stauffer D (ed) Annual Rev Comput Phys II. World Scientific, Singapore

Agent Based Modeling and Neoclassical Economics: A Critical Perspective*

SCOTT MOSS
Centre for Policy Modeling, Manchester Metropolitan University Business School, Manchester, UK

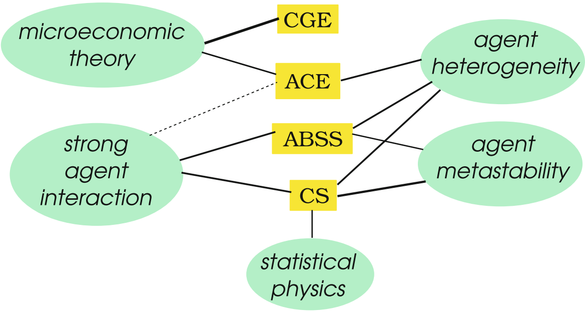
Article Outline

Introduction

Agent Based Modeling and Neoclassical Economics based modeling naturally generates complexity whereas neoclassical economics is incompatible in principle with complexity. The reasons that preclude complexity in neoclassical economic models also ensure that neoclassical economics cannot describe any society ever observed or that could ever be observed.

The meaning of complexity has been developed, mainly by physicists, to cover unpredictable, episodic volatility and also particular network topologies. In both cases there are nodes representing the components of a system and there are links among the components that can represent interactions amongst those components. Unpredictable, episodic volatility can result from particular forms of behavior by components and the interactions amongst those components. I am not aware of any investigations into relationships between that type of complexity and network topology.

The point to be made here is that the core assumptions *and* the methodology of conventional neoclassical economics preclude the emergence of episodic volatility and render social network topology inconsequential. When elaborated with heterogeneous agents, network topologies might have some effects on the outputs from computational neoclassical economic models – but, again, I am not aware of any systematic investigations into this possibility.



Agent Based Modeling and Neoclassical Economics: A Critical Perspective, Figure 1
Constraints on model designs

Economic Modeling Approaches

All definitions of complexity take for granted that there will be some specifications of individual components, that in general each component will interact with some other components and there will be some macro level phenomena that could not be described or understood except on the basis of the components and their interactions. The purpose of this section is to categorize the ways in which economists, agent based modelers and complexity scientists approach this micro-macro issue.

There are several strands in the economics and social sciences literatures for building macro analyzes explicitly on micro foundations. The main strands are computable general equilibrium (CGE), agent based computational economics (ACE), agent based social simulation (ABSS) and complexity science (CS) including econophysics and sociophysics. These strands are not all mutually exclusive although there are some conflicting elements among several of them.

Computable General Equilibrium

CGE is the most theoretically constrained of the four strands under consideration. As with general equilibrium theory, it is predicated on the assumptions that households maximize utility and firms maximize profits and that markets clear. The computational load associated with explicit representation of every household and firm leads to the adoption of representative agents intended to capture the behavior of a group such as all households or firms in a particular industrial sector. Some CGE models represent technology with input-output tables; others with marginalist production functions.

Agent Based Computational Economics

An objection to the representative agent device is raised in the ACE literature where the effects of agent hetero-

*The remarks about neoclassical economics are drawn from my inaugural lecture [28] geneity are explored. In these models, households can differ in their utility functions (or at least the parameters of those functions) or agents can adopt different game theoretic strategies and firms can employ different production functions. The theoretical core of ACE is not essentially different from that of CGE, both relying on conventional economic theory.

Agent Based Social Simulation

Models reported in the ABSS literature are by and large not driven by traditional theoretical concerns. There is a very wide range in the degree of empirical content: many models being developed to explore "stylized facts", others driven by qualitative case studies. The latter are often validated against both qualitative micro level data provided by stakeholders and against macro level statistical data.

Complexity Science

Because neoclassical economic theory excludes social embeddedness, the social complexity research that could be relevant to a consideration of neoclassical economics must be concerned with unpredictable, episodic turbulence. The CS literature on financial markets was seminal and remains well known. The interest in financial markets goes back to Mandelbrot [25] who used financial market data both because it exhibits "outliers" in the relative change series and because of the fineness of the time grain of the price and volume data. A seminal article by Palmer et al. [35] reported a simulation model in which individuals were represented by an early form of software agent and which produced time series marked by the occasional episodes of volatility of the sort observed in real financial market data. However, similar unpredictable episodes of turbulence and volatility have emerged in models of the early post-Soviet Russian economy [31], domestic water consumption [8,15] and models of transactions in intermediated markets [29]. Fine grain data exhibiting the same patterns of volatility were found *subsequent to the original publication* of each model.

Relationships Among the Four Approaches

The common thread between CGE and ACE is their common reliance on longstanding economic concepts of utility, continuous production functions, profit maximization, the use of game theory et sic hoc omnes. The common thread between ABSS and complexity science is the importance of social interaction and specifications of individual behavior that are either more ad hoc or based on detailed qualitative evidence for specific cases.

Complex, as distinct from rational, agents' behavior is "sticky": it takes non-trivial events or social pressures to make them change. This is the social meaning of metastability. They are also socially embedded in the sense that they interact densely with other agents and are influenced by some of those other agents, generally speaking others who are most like themselves and who they have reason to like and respect [10]. The social difference between influence and imitation is the social equivalent of the physical difference between dissipative and non-dissipative processes. Of course, such influence is meaningless unless the agents differ in some way – they must be heterogeneous.

Methodological Issues

Neoclassical economic theory has no empirically based micro foundation. It has agents of two types: households that maximize utility and firms that maximize profits. Time and expectations are allowed to influence these maximization processes by substituting "expected utility" or "expected profits" for the realized magnitudes. In such models, agents (households or firms) act as if they know with certainty a population distribution of possible outcomes from their actions. In the terminology introduced by Knight [23], risk pertains when the agent knows the *frequency* distribution of past outcomes that, as in actuarial contexts, are expected with confidence to pertain in the future. When no such frequency distribution is known, then uncertainty prevails. In the sense of Knight (though the terminology gets muddled in the economics literature), the assumption that agents maximize expected utility or expected profits is tenable in conditions of risk but not in conditions of uncertainty. Moreover, it has long been known (with Nobel prizes awarded to Allais [4] and to Daniel Kahneman of Kahneman and Tversky [21] for the demonstrations) that individuals do not act as if they were maximizers of utility or expected utility. Nor is there any evidence that enterprises actually maximize profits. Many economists acknowledge that rationality is bounded and that we lack the cognitive capacity to absorb the required amount of information and then to process that information in order to identify some optimal decision or action. This has given rise to a variety of schools of economic thought such as evolutionary economics (Nelson and Winter [32] is the seminal work here) and Keynesian economics [22] being amongst the best known.

There is evidently a recognizable (and often recognized) divide between the behavioral assumptions of neoclassical economics on the one hand and, on the other hand, common observation, experimental observation (cf. [5]) and a host of business histories (the work of Chandler [12,13] and Penrose [37] being surely the most influential). The evidence shows that the assumptions of neoclassical economics are inaccurate descriptions of the behavior the theories and models are purported to represent. There are two classes of defense for these descriptively inaccurate assumptions. On is the *as-if* defense and the other is the *for-simplicity* defense. These are considered in turn.

The as-if defense was enunciated in several forms by Samuelson [39], Friedman [17] and Alchian [3] in the years around 1950. The details of the differences between Samuelson and Friedman are not germane here. Both argued that their purpose was to model aggregate economic entities such as markets or national economies and descriptively inaccurate assumptions at micro level are permissible provided that the models are descriptively accurate at macro level. Alchian's contribution was to propose a mechanism. He asserted that, at least in the case of firms, those that were more profitably would be more likely to survive than firms that were less profitable. Consequently, over time, more and more firms would approach more closely to the maximum of profits available to them so that, even if they did not actually seek to maximize profits, the surviving population of firms would be those that implicitly did actually maximize profits.

The as-if defense is in practice an equilibrium argument. Neoclassical economic models are solved for the simultaneous maximization of utility and profits by all agents – or, at least it is proved that such a solution exists. In such a configuration, no agent has any incentive to change its behavior so the equilibrium presumed to be stable in the small (that is, once reached it is maintained). There is no proof that any general equilibrium model with an arbitrary number of agents is stable in the large (that is, that any feasible solution is an attractor of the system as a whole). The Alchian form of the as-if defense does not take into account any effects of agent interaction or influence of any agent by any other agent. In empirical – that is to say, econometric – testing of neoclassical models, extreme events and observations are dismissed as outliers and removed from the data set being used for the testing or else their effect is encapsulated by specially constructed dummy variables.

The for-simplicity defense rests on the presumption that simpler models are always to be preferred to more complicated models and the achievement of simplicity justifies making assumptions about behavior and environment that are not justified by evidence. The author has for many years now justified this claim by choosing any arbitrary leading economics journal and searching the most recent issue for an assumption made "for simplicity". On every occasion, the assumption made "for simplicity" has turned out to be an assumption that changed the nature of an empirical problem being addressed so that it conformed to the requirements (such as convexity or absence of externalities) of the mathematical technique being applied to the analysis. Seven of the eleven papers in, at the time of writing, the most recent (November 2007) issue of the *Quarterly Journal of Economics* appealed to the value of simplicity or, in one case, tractability to justify assumptions or specifications that were not justified empirically. The direct quotations are:

> Tractability obviously dictated the use of a simple summary statistic of the distribution of legislator ideologies (see p. 1418 in [14]).
>
> The fact, established below, that deriving the properties of the seats-votes relationship requires consideration only of the properties of the univariate distribution of ζ as opposed to those of the bivariate distribution of σ and μ considerably simplifies the analysis (see p. 1480 in [9]).
>
> We make three key assumptions to simplify the analysis. First, we assume that all jobs last indefinitely once found (i. e., there is no subsequent job destruction). Second, anticipating our empirical findings, we assume that wages are exogenously fixed, eliminating reservation-wage choices. Third, we assume that utility is separable in consumption and search effort (see p. 1516 in [11]).
>
> A more conventional timing assumption in search models without savings is that search in period t leads to a job that begins in period $t+1$. Assuming that search in period t leads to a job in period t itself simplifies the analytic expressions . . . (see p. 1517 in [11]).
>
> For simplicity, we'll assume that $\beta_{\text{SAT}}\left(X_{i,s}\right)$, $\beta_{\text{TEST}}\left(X_{i,s}\right)$ and $\beta_{\text{OTH}}\left(X_{i,s}\right)$ are linear in $X_{i,s}$ and can thus be written .... We will further assume that the random utility component is independent and identically distributed (i. i. d.) from a type 1 extreme value distribution (see p. 1616 in [19]).
>
> For simplicity, all households represent two-earner married couples of the same age (see p. 1683 in [33]).
>
> For simplicity, the model assumes that the highest 35 years of earnings correspond to the ages between 25 and 59 (see p. 1685 in [33]).
>
> We follow Auerbach and Kotlikoff (1987) by measuring efficiency gains from social security privatization using an LSRA that compensates households

> that would otherwise lose from reform. To be clear, the LSRA is not being proposed as an actual government institution. Instead, it is simply a hypothetical mechanism that allows us to measure the standard Hicksian efficiency gains in general equilibrium associated with privatization (see p. 1687 in [33]).

> Assume for simplicity that these batch sizes are fixed for each product class .... Given these fixed batch sizes for the two classes of product, the firm maximizes profits by deciding how many production runs ... [to] undertake ... (see pp. 1731–1732 in [7]).

> We adopt a number of simplifying assumptions to focus on the main implications of this framework. First, we assume that the relationship between the firm and each manager is short-term. Second, when $x_{i,k} = z_{i,k}$, the manager obtains a private benefit. We assume that managers are credit-constrained and cannot compensate principals for these private benefits and that these private benefits are sufficiently large so that it is not profitable for the principal to utilize incentive contracts to induce managers to take the right action. These assumptions imply that delegation will lead to the implementation of the action that is preferred by the manager ... (see p. 1769 in [1]).

All but the first of these quotations are from theoretical papers and the "simplifications" enable the authors to produce equilibrium solutions to their models. No one has ever knowingly observed an equilibrium and in a world where not everything is convex due to (for example) economies of large scale production and where computational capacities limit cognitive abilities, in principle no equilibrium ever will be observed. Indeed, Radner [38] showed that a *necessary* condition for general equilibrium to exist is that all agents have unlimited computational capacities if trading takes place at a sequence of dates. In the core general equilibrium model, all transactions are agreed at a single moment for all time. The "simplifications" required to produce equilibrium models cannot therefore be justified on the basis of relevance to empirical observation. They also ensure that the models cannot capture complexity.

Conditions for Complexity

The social and behavioral conditions for complexity manifest as unpredictable, episodic volatility appears to be the following:

- Individuals behave in routine ways unless some nontrivial event or events or social pressure from other individuals induce them to change their behavior.
- Individuals interact with other individuals.
- Individuals influence but do not generally imitate one another.
- Interactions amongst individuals and individual behavior are not dominated by events that do not arise from that interaction and behavior.

These conditions were first noticed as a general phenomenon in physical models and articulated by Jensen [20] as metastability, dense interaction, dissipation, and coldness of the system, respectively. The phenomenon of unpredictable, clustered volatility in social models had previously been noticed as had its similarity to self organized criticality as described by statistical physicists starting with Bak et al. [6].

Complexity and Social Volatility

Volatility in social statistical time series and power law distributed cross sectional data have long been observed by statisticians and social scientists. Vilfredo Pareto [36] discovered that the personal distribution of income is power law distributed, a finding which has been replicated widely across countries and time. The same phenomenon is now well known to characterize word use [43] city sizes [44], firm sizes (including market shares) [41], distributions of links between internet sites [2] and a host of other cross sectional distributions. Where firm sizes and market shares are concerned, there have been strands in the industrial economics literature reporting models yielding that result. However, the observed results have not been explained by models in which households maximize utility and firms maximize profits. As Simon and Bonini [41] point out, some variant to Gibrat's Law (or the law of proportional effect), which states that the growth rate of individuals (say firms) is not correlated with individual size, will generate one highly skewed distribution or another and the particular distribution can be refined by an appropriate choice of the representation of the law.

These desired results also emerged from a series of models based on plausible or empirically based specifications of individual behavior and social interaction in agent based social simulation models. In capturing stakeholders' perceptions of the behavior and social interactions of relevant classes of individuals and also in relying on well validated propositions from social psychology and cognitive science, models were implemented that produced the sort of skewed distributions that we observe in practice. Episodic volatility followed from the metastability and social embeddedness of agents. In the nature of this process, most changes are relatively small in magnitude but a few changes are very large. This results in fat-tailed distributions of relative changes in variable values at macro level and also in cross sectional data as described by Gibrat's Law. In practice, the large relative changes tend to be bunched together in unpredictable episodes of volatility.

While these results arise naturally in evidence-driven ABSS models, they are not easily reconciled with neoclassical economic theory. As Krugman [24] (quoted by Eeckhout [16]) had it, "We have to say that the rank-size rule is a major embarrassment for economic theory: one of the strongest statistical relationships we know, lacking any clear basis in theory".

Complexity and Social Network Topology

Social network topologies can obviously have no meaning in a model comprised by representative agents. In ACE, ABSS and CS models, there will always be a network of social interaction. However, the nature of the interaction can be very different across the different approaches.

In ACE models, it is common to find social interaction taking the form of games – typically the Prisoners' Dilemma. A nice example of such a model is Tesfatsion's labor market model using McFadzean and Tesfatsions's [26,42] Trading Network Game. In Tesfatsions's labour market model, there is a fixed number of workers and a fixed number of employers identical in their total offers of labour and of employment, respectively. Each worker (resp. employer) ascribes a value of utility to an employment arrangement with any employer (resp. worker). The utility starts out at a some exogenous value and is then increased or reduced depending on the experience at each trading date. The experience is the combination of cooperation and defection by each party to the employment relation at each time step. The social network in this model

> … is represented in the form of a directed graph in which the vertices $V(E)$ of the graph represent the work suppliers and employers, the edges of the graph (directed arrows) represent work offers directed from work suppliers to employers, and the edge weight on any edge denotes the number of accepted work offers (contracts) between the work supplier and employer connected by the edge (see p. 431 in [42]).

The topology of this network depends on the outcomes of sequences of prisoners' dilemma games determining the utilities of workers and employers to one another. Every worker can see every employer and conversely so that the directed links between agents are limited by the number of work contracts into which each agent can engage. After some arbitrary number of time steps, the strategies of the agents are represented as genes and a genetic algorithm is applied so that, over a whole simulation, the elements of the most successful defect/cooperate strategies become more dominant. Since these strategies determine the outcomes of the prisoners' dilemma games, the social network continues to evolve with utility enhancing strategies becoming more dominant.

In a recent (at the time of writing) issue of The Journal of Economic Dynamics and Control, Page and Tassier [34] modeled the development of chain stores across markets. A firm was defined by its product. Each product was assigned an "intrinsic quality" represented by an integer drawn at random from a distribution $\theta(q)$, and a set of I "hedonic attributes" represented by I positive integers in a range from 0 to some arbitrary, user selected number A. Consumers are represented by utility functions that are positively related to "quality" and negatively related to the difference between some desired set of hedonic attributes and the hedonic attributes of the product. There are a number (set by the model user) of discrete markets. Page and Tassier then run a variety of simulations that allow for firms to replicate themselves across markets or, through lack of demand, to leave markets.

These two models seem to be representative of a wide class of ACE models. In the first place, agents are defined by utility functions or game theoretic strategies so that the behavior of any individual agent is either fixed or responds smoothly to infinitesimal changes in prices, incomes or whatever other arguments might populate its utility function. In either event, agents cannot be metastable and follow behavioral routines until (but only until) some significant stimulus causes them to change their behavioral responses. In the second place, agents' preferences and responses are not influenced by the preferences or actions of any other agents like themselves. That is, their behavior as determined by their utility functions or game theoretic strategies will respond to market signals or the actions of the other agent in their game but not to communications with or observations of any other agents. These agents are not, in the words of Granovetter [18], socially embedded especially since it is rare in a neoclassical model for there to be more than two players in a game and unheard-of for there to be more than three (cf. [27]).

Whilst we cannot state with authority that the conditions of metastability, social influence and the consistency principle are necessary for complexity to emerge at macro level from micro level behavior, these conditions have characterized the social simulation models that have produced the episodic and unpredictable volatility associated with complexity. The absence of social embeddedness in the neoclassical ACE models must also explain their lack of any representation of social (as distinct from merely economic) networks.

Complexity and the Role of Evidence

An interesting and fairly typical feature of papers reporting neoclassical models – both theoretical and computational with agents – is that they motivate the modeling exercise by appeal to some empirical, macro level economic phenomenon and then ignore evidence about the micro level behavior that might bring about such phenomena. This practice can be seen in both of the ACE examples described in Sect. "Conditions for Complexity".

Tesfatsion [42] motivates her model on more theoretical grounds than do Page and Tassier [34]. She wrote:

> Understanding the relationship between market structure, market behavior, and market power in markets with multiple agents engaged in repeated strategic interactions has been a major focus of analytical, empirical, and human-subject experimental researchers in industrial organization since the early 1970s. To date, however, definitive conclusions have been difficult to obtain.

She goes on to cite "a unified theoretical treatment of oligopoly decision-making", an article on empirical findings with respect to market power that looks only at industry level statistical measures, and some work with experimental subjects. No references are made, either by Tesfatsion or those she cites, to any case studies of the "repeated strategic interactions" in which the "multiple agents" engage.

Page and Tassier give more historical detail. Their motivation turns on:

> Chain stores and franchises dominate the American economic landscape. A drive through any moderately sized city reveals remarkable conformity in restaurants, stores, and service centers. Anyone who so desires can eat at Applebee's, shop at Wal-Mart, and grab a Starbuck's latte grande while getting her car brakes done at Midas (see p. 3428 in [34]).
>
> For example, in many markets, Lowe's and Home Depot capture a significant portion of the home improvement market. These big box stores drove many small independent hardware stores and lumber yards out of business. The residual demand resides in niches that can be filled by hardware chains specializing in upscale home furnishings like Restoration Hardware. … Often, when Lowe's enters a market, it creates a niche for Restoration Hardware as well. And, as both Lowe's and Restoration Hardware enter more and more markets, they in turn create additional common niches that can be filled by even more chains. Thus, chains beget chains (see p. 3429 in [34]).

So this article claims a clear and direct historical basis. And yet

> … To capture the increasing correlation in niches formally, we introduce two new concepts, the niche landscape and the differential niche landscape. The former plots the quality required to enter a market at a given set of hedonic attributes. The latter plots the differences in two niche landscapes. In the presence of chains, differential niche landscapes become flat, i. e. the niche landscapes become correlated across markets (see p. 3429 in [34]).

The representation of the actors in this framework has been discussed above. At no stage is the agent design discussed in relation to any empirical accounts of the behavior and motivations of the managers of Wal-Mart, Starbucks, Lowes, Restoration Hardware or any other enterprise or any consumer.

This is, of course, the way of neoclassical economics and it has extended to ACE research as well. What is perhaps more unsettling is that it has also extended to the bastions of complexity science – the econophysicists.

There is a long literature now on complexity and financial markets and also on complexity an the formation of opinions – opinion dynamics. There are at least two reasons for the popularity amongst physicists of financial market modeling. First, there are long series of very fine grain data. Second, the data exhibits the unpredictable, episodic volatility associated with complexity. The popularity of opinion dynamics cannot be based on the quality of the data – even at macro level – because that quality is much more coarse grain and inconsistent over time than financial market data. Nonetheless, the two literatures are marked by the heavy presence and influence of physicists and by the lack of connection between their agent designs and any available evidence about the behavior of traders in financial markets or voters or others acting on or expressing their opinions.

A good example from the opinion dynamics literature – chosen at random from *The European Physical Journal B* – is by Schweitzer and Hoyst [40], "Modeling collective opinion formation by means of active Brownian particles". The motivation for their article is

> The formation of public opinion is among the challenging problems in social science, because it reveals a complex dynamics, which may depend on different internal and external influences. We mention the influence of political leaders, the biasing effect of mass media, as well as individual features, such as persuasion or support for other opinions.

We immediately have complexity and social relevance to motivate an article on social dynamics in a physics journal. However, there is no empirical justification for modeling individuals who form opinions as active Brownian particles. The apparent complexity in the outcomes of social processes of opinion formation can be produced by the non-linear feedbacks of fields of active Brownian particles. Whether individuals actually behave in this way is not addressed by Schweitzer and Hoyst or, as far as I know, by any contributor to the opinion dynamics literature.

Much the same can be said of the econophysics literature on financial markets. The clustered volatility associated with complexity is readily produced by physical models with characteristics of metastability, dissipation and dense patterns of interaction. What the econophysicists fail to address is the question of whether their particular formulations – and active Brownian particle is just one of many examples – are *descriptively accurate* representations of the individual actors whose behavior they are seeking to analyze.

In this regard, the econophysicists are not better scientists than neoclassical economists. It can be said in favor of neoclassical (including ACE) economists that they are at least following in a long tradition when they ignore the relationship between what people actually do and how agents are modeled. In the long history of the physical sciences, however, observation and evidence at micro and macro level and all levels in between has dominated theory (cf. [30]). There are some at least in the ABSS research community who would prefer our colleagues with backgrounds in the physical scientists to follow their own methodological tradition in this regard and not that of the neoclassical economists.

Future Directions

Complexity science is not a niche research interest in the social sciences. Societies are complex and all social science should be complexity science. However, any social science that excludes social interaction and inertia or routine necessarily suppresses complexity. As noted here, the adoption of utility theory and representative agents by neoclassical economists (and other social scientists influenced by them) amounts to the exclusion of behavioral inertia and social interaction, respectively. To drop both utility and representative agents and to build analyzes bottom up from a sound basis in evidence would produce a better – very likely, a good – body of economic analysis. But the transition from present convention would be enormous – a transition that experience shows to be beyond the capacity of current and previous generations of mainstream economists. Not only would they have to abandon theories that drive and constrain their research but also their whole epistemological and wider methodological stance. They would have to accept that prediction and forecasting cannot be core methodological objectives and that theories are built by abstracting from detailed evidence based social simulation models the designs and outputs from which have been validated by stakeholders in a range of contexts. This would be a future direction guided by good science.

Bibliography

1. Acemoglu D, Aghion P, Lelarge C, Van Reenen J, Zilibotti F (2007) Technology, information, and the decentralization of the firm. Q J Econ 122(4):1759–1799. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1759
2. Adamic LA, Huberman BA (1999) Power-law distribution of the World Wide Web. Science 287(5461):2115
3. Alchian AA (1950) Uncertainty, evolution and economic theory. J Political Econ 58(2):211–221
4. Allais M (1953) Le comportement de l'homme rationnel devant le risque: Critiques des postulats et axiomes de l'ecole americaine. Econometrica 21(4):503–546
5. Anderson JR (1993) Rules of the mind. Lawrence Erlbaum Associates, Hillsdale
6. Bak P, Tang C, Weisenfeld K (1987) Self organized criticality: An explanation of 1/f noise. Phys Rev Lett 59(4):381–384
7. Bartel A, Ichniowski C, Shaw K (2007) How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. Q J Econ 122(4):1721–1758. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1721
8. Barthelemy O (2006) Untangling scenario components with agent based modeling: An example of social simulations of water demand forecasts. Ph D thesis, Manchester Metropolitan University
9. Besley T, Preston I (2007) Electoral bias and policy choice: Theory and evidence. Q J Econ 122(4):1473–1510. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1473
10. Brown R (1965) Social psychology. The Free Press, New York
11. Card D, Chetty R, Weber A (2007) Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. Q J Econ 122(4):1511–1560. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1511

12. Chandler AD (1962) Strategy and structure: Chapters in the history of the american industrial enterprise. MIT Press, Cambridge
13. Chandler AD (1977) The visible hand: The managerial revolution in american business. Harvard University Press
14. Coate S, Brian K (2007) Socially optimal districting: A theoretical and empirical exploration. Q J Econ 122(4):1409–1471
15. Downing TE, Moss S, Pahl Wostl C (2000) Understanding climate policy using participatory agent based social simulation. In: Moss S, Davidsson P (eds) Multi agent based social simulation. Lecture Notes in Artificial Intelligence, vol 1979. Springer, Berlin, pp 198–213
16. Eeckhout J (2004) Gibrat's law for (all) cities. Am Econ Rev 94(5):1429–1451
17. Friedman M (1953) The methodology of positive economics. In: Essays on Positive Economics. University of Chicago Press, Chicago
18. Granovetter M (1985) Economic action and social structure: The problem of embeddedness. Am J Sociol 91(3):481–510
19. Jacob BA, Lefgren L (2007) What do parents value in education? An empirical investigation of parents' revealed preferences for teachers. Q J Econ 122(4):1603–1637. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1603
20. Jensen H (1998) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
21. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47(2):263–292
22. Keynes JM (1935) The general theory of employment, interest and money. Macmillan, London
23. Knight FH (1921) Risk, uncertainty and profit. Houghton-Mifflin
24. Krugman P (1995) Development, geography and economic theory. MIT Press, Cambridge
25. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36(4):394–419
26. McFadzean D, Tesfatsion L (1999) A c++ platform for the evolution of trade networks. Comput Econ 14:109–134
27. Moss S (2001) Game theory: Limitations and an alternative. J Artif Soc and Soc Simul 4(2)
28. Moss S (1999) Relevance, realism and rigour: A third way for social and economic research. Technical Report 99-56, Centre for Policy Modeling, Manchester Metropolitan University
29. Moss S (2002) Policy analysis from first principles. Proc US Nat Acad Sci 99(Suppl. 3):7267–7274
30. Moss S, Edmonds B (2005) Towards good social science. J Artif Soc and Soc Simul 8(4). ISSN 1460-7425. http://jasss.soc.surrey.ac.uk/8/4/13.html
31. Moss S, Kuznetsova O (1996) Modeling the process of market emergence. In: Owsinski JW, Nahorski Z (eds) Modeling and analyzing economies in transition. MODEST, Warsaw, pp 125–138
32. Nelson RR, Winter SG (1982) An evolutionary theory of economic change. Harvard University Press, Cambridge
33. Nishiyama S, Smetters K (2007) Does social security privatization produce efficiency gains? Q J Econ 122(4):1677–1719. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1677
34. Page SE, Tassier T (2007) Why chains beget chains: An ecological model of firm entry and exit and the evolution of market similarity. J Econ Dyn Control 31(10):3427–3458
35. Palmer R, Arthur WB, Holland JH, LeBaron B, Taylor P (1993) Artificial economic life: A simple model for a stock market. Physica D 75:264–274
36. Pareto V (1896–1897) Cours d'économie politique professé à l'Université de Lausanne. Rouge, Lausanne
37. Penrose ET (1959) The theory of the growth of the firm. Wiley, New York
38. Radner R (1968) Competitive equilibrium under uncertainty. Econometrica 36(1):31–58
39. Samuelson PA (1949) Foundations of economic analysis. Harvard University Press
40. Schweitzer F, Hoyst JA (2000) Modeling collective opinion formation by means of active Brownian particles. Europ Phys J B – Condens Matter Complex Syst 15(4):723–732
41. Simon HA, Bonini CP (1958) The size distribution of business firms. Am Econ Rev 48(4):607–617. ISSN 0002-8282 http://links.jstor.org/sici?sici=0002-8282%28195809%2948%3A4%3C607%3ATS%DOBF%3E2.0.CO%3B2-3
42. Tesfatsion L (2001) Structure, behavior, and market power in an evolutionary labor market with adaptive search. J Econ Dyn Control 25(3–4):419–457
43. Zipf GK (1935) The psycho-biology of language. Houghton Mifflin, Boston
44. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Cambridge

Agent Based Models in Economics and Complexity

MAURO GALLEGATI[1], MATTEO G. RICHIARDI[1,2]
[1] Università Politecnica delle Marche, Ancona, Italy
[2] Collegio Carlo Alberto – LABORatorio R. Revelli, Moncalieri, Italy

Article Outline

Glossary

Abduction also called *inference to the best explanation*, abduction is a method of reasoning in which one looks for the hypothesis that would best explain the relevant evidence.

Agents entities of a model that (i) are perceived as a unit from the outside, (ii) have the ability to *act*, and possibly to *react* to external stimuli and *interact* with the environment and other agents.

Agent-based computational economics (ACE) is the computational study of economic processes modeled as dynamic systems of interacting agent.

Agent-based models (ABM) are models where (i) there is a multitude of objects that interact with each other and with the environment; (ii) the objects are autonomous, i. e. there is no central, or top-down control over their behavior; and (iii) the outcome of their interaction is numerically computed.

Complexity there are more than 45 existing definitions of complexity (Seth Lloyd, as reported on p. 303 in [97]). However, they can be grouped in just two broad classes: a *computational* view and a *descriptive* view. Computational (or algorithmic) complexity is a measure of the amount of information necessary to compute a system; descriptive complexity refers to the amount of information necessary to describe a system. We refer to this second view, and define complex systems as systems characterized by emergent properties (see *emergence*).

Deduction the logical derivation of conclusions from given premises.

Economics is the science about the intended and unintended consequences of individual actions, in an environment characterized by scarce resources that both requires and forces to interaction.

Emergence the spontaneous formation of self-organized structures at different layers of a hierarchical system configuration.

Evolution in biology, is a change in the inherited traits of a population from one generation to the next. In social sciences it is intended as an endogenous change over time in the behavior of the population, originated by competitive pressure and/or learning.

Heterogeneity non-degenerate distribution of characteristics in a population of agents.

Induction the intuition of general patterns from the observation of statistical regularities.

Interaction a situation when the actions or the supposed actions of one agent may affect those of other agents within a reference group.

Out-of-equilibrium a situation when the behavior of a system, in terms of individual strategies or aggregate outcomes, is not stable.

Definition of the Subject

A crucial aspect of the complexity approach is how interacting elements produce aggregate patterns that those elements in turn react to. This leads to the emergence of aggregate properties and structures that cannot be guessed by looking only at individual behavior.

It has been argued [144] that complexity is ubiquitous in economic problems (although this is rarely acknowledged in economic modeling), since (i) the economy is inherently characterized by the interaction of individuals, and (ii) these individuals have cognitive abilities, e. g. they form expectations on aggregate outcomes and base their behavior upon them: "Imagine how hard physics would be if electrons could think", is how the Nobel prize winner Murray Gell–Mann, a physicist, has put it (as reported by Page [131]).

Explicitly considering how heterogeneous elements dynamically develop their behavior through interaction is a hard task analytically, the equilibrium analysis of mainstream (neoclassical) economics being a shortcut that in many cases is at risk of throwing the baby out with the bath water, so to speak. On the other hand, numerical computation of the dynamics of the process started to be a feasible alternative only when computer power became widely accessible. The computational study of heterogeneous interaction agents is called agent-based modeling (ABM). Interestingly, among its first applications a prominent role was given to economic models [4], although it was quickly found of value in other disciplines too (from sociology to ecology, from biology to medicine). The goal of this chapter is to motivate the use of the complexity approach and agent-based modeling in economics, by discussing the weaknesses of the traditional paradigm of mainstream economics, and then explain what ABM is and which research and policy questions it can help to analyze.

Introduction

Economics is in troubled waters. Although there exists a mainstream approach, its internal coherence and ability to explain the empirical evidence are increasingly questioned. The causes of the present state of affairs go back to the middle of the eighteenth century, when some of the Western economies were transformed by the technological progress which lead to the industrial revolution. This was one century after the Newtonian revolution in physics: from the small apple to the enormous planets, all objects seemed to obey the simple *natural* law of gravitation. It was therefore natural for a new figure of social scientist, the economist, to borrow the method (*mathematics*) of the most successful hard science, physics, allowing for the mutation of *political economy* into *economics*. It was (and still is) the mechanical physics of the seventeenth century, which ruled economics. In the final chapter of his General Theory, Keynes wrote of politicians as slaves of late economists: in their turn, they are slaves of late physicists of the seventeenth century (see also [125]).

From then on, economics lived its own evolution based on the classical physics assumptions (reductionism, determinism and mechanicism). Quite remarkably, the approach of statistical physics, which deeply affected physical science at the turn of the nineteenth century by emphasizing the difference between micro and macro, was adopted by Keynes around the mid 1930s. However, after decades of extraordinary success it was rejected by the neoclassical school around the mid 1970s, which framed the discipline into the old approach and ignored, by definition, any interdependencies among agents and difference between individual and aggregate behavior (being agents, electrons, nations or planets).

The ideas of natural laws and equilibrium have been transplanted into economics *sic et simpliciter*. As a consequence of the adoption of the classical mechanics paradigm, the difference between micro and macro was analyzed under a reductionist approach. In such a setting, aggregation is simply the process of summing up market outcomes of individual entities to obtain economy-wide totals. This means that there is no difference between micro and macro: the dynamics of the whole is nothing but a summation of the dynamics of its components (in term of physics, the motion of a planet can be described by the dynamics of the atoms composing it). This approach does not take into consideration that there might be two-way interdependencies between the agents and the aggregate properties of the system: interacting elements produce aggregate patterns that those elements in turn react to. What macroeconomists typically fail to realize is that the correct procedure of aggregation is not a sum: this is when emergence enters the drama. With the term emergence we mean the arising of complex structures from simple individual rules [147,153,171]. Empirical evidence, as well as experimental tests, shows that aggregation generates regularities, i. e. simple individual rules, when aggregated, produce statistical regularities or well-shaped aggregate functions: regularities emerge from individual chaos [106]. The concept of equilibrium is quite a dramatic example. In many economic models equilibrium is described as a state in which (individual and aggregate) demand equals supply. The notion of *statistical equilibrium*, in which the aggregate equilibrium is compatible with individual disequilibrium, is outside the box of tools of the mainstream economist. The same is true for the notion of *evolutionary equilibrium* (at an aggregate level) developed in biology. The equilibrium of a system no longer requires that every single element be in equilibrium by itself, but rather that the statistical distributions describing aggregate phenomena be stable, i. e. in "[...] a state of macroscopic equilibrium maintained by a large number of transitions in opposite directions" (p. 356 in [64]).

According to this view, an individual organism is in equilibrium *only when it is dead*. A consequence of the idea that macroscopic phenomena can emerge is that reductionism is wrong.

Ironically, since it can be argued, as we will do in the section below, that economics strongly needs this methodological twist [144], ABM has received lees attention in economics than in other sciences ([110]; but [82] is a counter-example). The aim of this chapter is not to provide a review of applications of the complexity theory to economics (the interested reader is referred to [15,26,60,124,140,142]), but rather to describe the development of the *Agent-Based Modeling* (ABM) approach to complexity.

The chapter is structured as follows: after reviewing some limits of mainstream economics (Sect. "Additional Features of Agent-Based Models"), Sects. "The Economics of Complexity" and "Additional Features of Agent-Based Models" describe how the complexity perspective differs from the traditional one, and how many problems of the mainstream approach can be overcome by ABM. As an example, we present a prototypical example of ABM, based on the work of Thomas Schelling on the dynamics of segregation. After dedicating some sections to, respectively, a skeleton history of ABM, a recursive system representation of these models, a discussion on how ABM can be interpreted, estimated and validated, we finally discus how the complexity approach can be used to guide policy intervention and analysis. A final section discusses the achievements of the ABM agenda.

Some Limits of the Mainstream Approach

The research program launched by the neoclassical school states that macroeconomics should be explicitly grounded on microfoundations. This is how Robert Lucas put it: "The most interesting recent developments in macroeconomic theory seem to me describable as the reincorporation of aggregative problems [...] within the general framework of 'microeconomic' theory. If these developments succeed, the term 'macroeconomic' will be simply disappear from use and the modifier 'micro' will become superfluous. We will simply speak, as did Smith, Marshall and Walras, of *economic theory*" (pp. 107–108 in [115]). According to the mainstream, this implies that economic phenomena at a macroscopic level should be explained as a summation of the activities undertaken by individual decision makers. This procedure of microfoundation is very different from that now used in physics. The latter starts from the micro-dynamics of the single particle, as expressed by the Liouville equation and, through the master equation, ends up with the macroscopic equations. In the aggregation process, the dynamics of the agents lose their degree of freedom and behave coherently in the aggregate. In mainstream economics, while the procedure is formally the same (from micro to macro), it is assumed that the dynamics of the agents are those of the aggregate. The reduction of the degree of freedom, which is characteristic of the aggregation problem in physics, is therefore ruled out: a rational agent with complete information chooses to implement the individually optimal behavior, without additional constraints. There are three main pillars of this approach: (i) the precepts of the rational choice-theoretic tradition; (ii) the equilibrium concept of the Walrasian analysis; and (iii) the reductionist approach of classical physics. In the following, we will show that assumptions (i)–(ii), which constitute the necessary conditions for reducing macro to micro, are logically flawed (and empirically unfounded), while rejection of (iii) opens the road to complexity.

Mainstream economics is axiomatic and based on unrealistic (or unverifiable) assumptions. According to the supporters of this view, such an abstraction is necessary since the real world is complicated: rather than compromising the epistemic worth of economics, such assumptions are essential for economic knowledge. However, this argument does not invalidate the criticism of unrealistic assumptions [136]. While it requires internal coherence, so that theorems can be logically deduced from a set of assumptions, it abstracts from external coherence between theoretical statements and empirical evidence. Of course, this implies an important epistemological detachment from falsifiable sciences like physics. In setting the methodological stage for the dynamic stochastic general equilibrium (DSGE) macroeconomic theory, Lucas and Sargent declared:

> "An economy following a multivariate stochastic process is now routinely described as being in equilibrium, by which is meant nothing more that at each point in time (a) markets clears and (b) agents act in their own self-interest. This development, which stemmed mainly from the work of Arrow [...] and Debreu [...], implies that simply to look at any economic time series and conclude that it is a disequilibrium phenomenon is a meaningless observation. [...] The key elements of these models are that agents are rational, reacting to policy changes in a way which is in their best interests privately, and that the impulses which trigger business fluctuations are mainly unanticipated shocks." (p. 7 in [116]).

The self-regulating order of Adam Smith [153] is transformed into a competitive general equilibrium (GE) in the form elaborated in the 1870s by Walras, that is a configuration of (fully flexible) prices and plans of action such that, at those prices, all agents can carry out their chosen plans and, consequently, markets clear. In a continuous effort of generalization and analytical sophistication, modern (neoclassical) economists interested in building microfoundations for macroeconomics soon recurred to the refinement proposed in the 1950s by Arrow and Debreu [14], who showed that also individual intertemporal (on an infinite horizon) optimization yields a GE, as soon as the economy is equipped with perfect price foresight for each future state of nature and a complete set of Arrow-securities markets [11], all open at time zero and closed simultaneously. Whenever these conditions hold true, the GE is an allocation that maximizes a properly defined social welfare function, or the equilibrium is Pareto-efficient (*First Welfare Theorem*).

The literature has pointed out several logical inconsistencies of the mainstream approach. Davis [44] identifies three impossibility results, which determine the breakdown of the mainstream, i. e. neoclassical, economics: (i) Arrow's 1951 theorem showing that neoclassical theory is unable to explain social choice [10]; (ii) the Cambridge capital debate pointing out that mainstream is contradictory with respect to the concept of aggregate capital [40]; and (iii) the Sonnenschein–Mantel–Debreu results showing that the standard comparative static reasoning is inapplicable in general equilibrium models. In particular, a few points are worth remembering here.

1. The GE is neither unique nor locally stable under general conditions. This negative result, which refers to the work of Sonnenschein [155], Debreu [46] and Mantel [119], can be summarized along the following lines. Let the aggregate excess demand function $F(p)$ – obtained from aggregating among individual excess demands $f(p)$ – be a mapping from the price simplex Π to the commodity space P^N. A GE is defined as a price vector p such that $F(p*) = 0$. It turns out that the only conditions that $F(\cdot)$ inherits from $f(\cdot)$ are continuity, homogeneity of degree zero and the Walras' law (i. e., the total value of excess demand is zero). These assure the existence, but neither the uniqueness nor the local stability of $p*$, unless preferences generating individual demand functions are restricted to very implausible cases.
2. The existence of a GE is proved *via* the Brower's fix point theorem, i. e. by finding a continuous function $g(\cdot)\colon \Pi \rightarrow \Pi$ so that any fixed point for $g(\cdot)$ is also an equilibrium price vector $F(p*) = 0$. Suppose that we are interested in finding an algorithm which, starting from an arbitrary price vector p, chooses price sequences to check for $p*$ and halts when it finds it. In other terms, to find the GE price vector $F(p*) = 0$ means that halting configurations are decidable. As this violates the undecidability of the halting problem for Turing machines, from a recursion theoretic viewpoint the GE solution is incomputable [138,167]. Notice that the same problem applies, in spite of its name, to the class of computable GE models [169].
3. By construction, in a GE all transactions are undertaken at the same equilibrium price vector. Economic theory has worked out two mechanisms capable of reaching this outcome. First, one can assume that buyers and sellers adjust, costless, their optimal supplies and demands to prices called out by a (explicit or implicit) fictitious auctioneer, who continues to do his job until he finds a price vector which clears all markets. Only then transactions take place (Walras' assumption). Alternatively, buyers and sellers sign provisional contracts and are allowed to freely (i. e., without any cost) recontract until a price vector is found which makes individual plans fully compatible. Once again, transactions occur only after the equilibrium price vector has been established (Edgeworth's assumption). Regardless of the mechanism one adopts, the GE model is one in which the formation of prices precedes the process of exchange, instead of being the result of it, through a *tatonnement* process occurring in a meta-time. Real markets work the other way round and operates in real time, so that the GE model cannot be considered a scientific explanation of real economic phenomena [9].
4. It has been widely recognized since Debreu [45], that integrating money in the theory of value represented by the GE model is at best problematic. No economic agent can individually decide to monetize alone; monetary trade should be the equilibrium outcome of market interactions among optimizing agents. The use of money – that is, a common medium of exchange and a store of value – implies that one party to a transaction gives up something valuable (for instance, his endowment or production) for something inherently useless (a fiduciary token for which he has no immediate use) in the hope of advantageously re-trading it in the future. Given that in a GE model actual transactions take place only after a price vector coordinating all trading plans has been freely found, money can be consistently introduced into the picture only if the logical keystone of the absence of transaction costs is abandoned. By the same token, since credit makes sense only if agents can sign contracts in which one side promises future delivery of goods or services to the other side, in equilibrium markets for debt are meaningless, and bankruptcy can be safely ignored. Finally, as the very notion of a GE implies that all transactions occur only when individual plans are mutually compatible, and this has to be true also in the labor market, the empirically observed phenomenon of involuntary unemployment and the microfoundation program put forth by Lucas and Sargent are logically inconsistent.

5. The very absence of money and credit is a consequence of the fact that in GE there is no time. The only role assigned to time in a GE model is, in fact, that of dating commodities. Products, technologies and preferences are exogenously given and fixed from the outset. The convenient implication of banning out-of-equilibrium transactions is simply that of getting rid of any disturbing influence of intermediary modifications of endowments – and therefore of individual excess demands – on the final equilibrium outcome. The introduction of non-Walrasian elements into the GE microfoundations program – such as fixed or sticky prices, imperfect competition and incomplete markets leading to temporary equilibrium models – yields interesting Keynesian features such as the breaking of the Say's law and scope for a monetary theory of production, a rationale for financial institutions and a more persuasive treatment of informational frictions. As argued in Vriend [165], however, all these approaches preserve a Walrasian perspective in that models are invariably closed by a GE solution concept which, implicitly or (more often) not, implies the existence of a fictitious auctioneer who processes information, calculates equilibrium prices and quantities, and regulates transactions. As a result, if the Walrasian auctioneer is removed the decentralized economy becomes dynamically incomplete, as we are not left with any mechanism determining how quantities and prices are set and how exchanges occur.

The flaws of the solution adopted by mainstream macroeconomists to overcome the problems of uniqueness and stability of equilibrium on the one hand, and of analytical-tractability on the other one – i. e. the usage of a representative agent (RA) whose choices summarize those of the whole population of agents – are so pervasive that we discuss them hereafter.

6. Although the RA framework has a long history, it is standard to build the microfoundation procedure on it only after Lucas' critique paper [114]. Mainstream models are characterized by an explicitly stated optimization problem of the RA, while the derived individual demand or supply curves are used to obtain the aggregate demand or supply curves. Even when the models allow for heterogeneity, interaction is generally absent (the so-called *weak interaction hypothesis* [139]). The use of RA models should allow one to avoid the Lucas critique, to provide microfoundations to macroeconomics, and, *ça va sans dire*, to build Walrasian general equilibrium models. Since models with many heterogeneous interacting agents are complicated and no closed form solution is often available (aggregation of heterogenous interacting agents is analyzed in [5,6,7,53,78]), economists assume the existence of an RA: a simplification that makes it easier to solve for the competitive equilibrium allocation, since direct interaction is ruled out by definitions. Unfortunately, as Hildenbrand and Kirman [95] noted:

> "There are no assumptions on isolated individuals, which will give us the properties of aggregate behavior. We are reduced to making assumptions at the aggregate level, which cannot be justified, by the usual individualistic assumptions. This problem is usually avoided in the macroeconomic literature by assuming that the economy behaves like an individual. Such an assumption cannot be justified in the context of the standard model".

The equilibria of general equilibrium models with a RA are characterized by a complete absence of trade and exchange, which is a counterfactual idea. Kirman [99], Gallegati [76] and Caballero [36] show that RA models ignore valid aggregation concerns, by neglecting interaction and emergence, hence committing fallacy of composition (what in philosophy is called fallacy of division, i. e. to attribute properties to a different level than where the property is observed: game theory offers a good case in point with the concept of Nash equilibrium, by assuming that social regularities come from the agent level equilibrium). Those authors provide examples in which the RA does not represent the individuals in the economy so that the reduction of a group of heterogeneous agents to an RA is not just an analytical convenience, but it is both unjustified and leads to conclusions which are usually misleading and often wrong ([99]; see also [98]). A further result, which is a proof of the logical fallacy in bridging the micro to the macro is the *impossibility theorem* of Arrow: it shows that an ensemble of people, which has to collectively take a decision, cannot show the same rationality of an individual [123]. Moreover, the standard econometric tools are based upon the assumption of an RA. If the economic system is populated by heterogeneous (not necessarily interacting) agents, then the problem of the microfoundation of macroeconometrics becomes a central topic, since some issues (e. g., *co-integration*, *Granger-causality*, *impulse-response function of structural VAR*) lose their significance [69].

All in all, we might say that the failure of the RA framework, points out the *vacuum* of the mainstream microfoundation literature, which ignores interactions: no box of tools is available to connect the micro and the macro levels, beside the RA whose existence is at odds with the empirical evidence [30,158] and the equilibrium theory as well [99].

The Economics of Complexity

According to the mainstream approach there is no direct interaction among economic units (for a pioneeristic and neglected contribution see [68]; see also [101]). In the most extreme case, any individual strategy is excluded (*principle of excluded strategy*, according to Schumpeter [149]) and agents are homogeneous. Small departures from the perfect information hypothesis are incoherent with the Arrow–Debreu general equilibrium model, as shown by Grossman and Stiglitz [88], since they open the chance of having direct links among agents [156]. In particular, if prices convey information about the quality there cannot be an equilibrium price as determined by the demand-supply schedule, since demand curves depend on the probability distribution of the supply (p. 98 in [87]).

What characterizes a complex system is the notion of emergence, that is the spontaneous formation of self-organized structures at different layers of a hierarchical system configuration [43]. Rather, mainstream economics conceptualizes economic systems as consisting of several identical and isolated components, each one being a copy of a RA. The aggregate solution can thus be obtained by means of a simple summation of the choices made by each optimizing agent. The RA device, of course, is a way of avoiding the problem of aggregation by eliminating heterogeneity. But heterogeneity is still there. If the macroeconomist takes it seriously, he/she has to derive aggregate quantities and their relationships from the analysis of the micro-behavior of different agents. This is exactly the key point of the *aggregation problem*: starting from the *micro-equations* describing/representing the (optimal) choices of the economic units, what can we say about the *macro-equations*? Do they have the same functional form of the micro-equations (the *analogy principle*)? If not, how is the macro-theory derived?

The complexity approach to economics discards the GE approach to the microfoundation program, as well as its RA shorthand version. Instead of asking to deductively prove the existence of an equilibrium price vector $p*$ such that $F(p*) = 0$, it aims at explicitly constructing it by means of an algorithm or a rule. From an epistemological perspective, this implies a shift from the realm of classical to that of constructive theorizing [168]. Clearly, the act of computationally constructing a coordinated state – instead of imposing it via the Walrasian auctioneer – for a decentralized economic system requires complete description of goal-directed economic agents and their interaction structure.

Agent-based modeling represents an effective implementation of this research agenda ([60,124], see also [24,67,81,175]). ABM is a methodology that allows one to construct, based on simple rules of behavior and interaction, models with heterogeneous agents, where the resulting aggregate dynamics and empirical regularities are not known a priori and are not deducible from individual behavior. It is characterized by three main tenets: (i) there is a multitude of objects that interact with each other and with the environment; (ii) the objects are autonomous, i. e. there is no central, or top-down control over their behavior; and (iii) the outcome of their interaction is numerically computed. Since the objects are autonomous, they are called agents ([3,4]; see also the repository of ACE-related material maintained by Leigh Tesfatsion at http://www.econ.iastate.edu/tesfatsi/ace.htm): "Agent-based Computational Economics is the computational study of economic processes modeled as dynamic systems of interacting agent" [161].

Agents can be anything from cells to biological entities, from individuals to social groups like families or firms. Agents can be composed by other agents: the only requirement being that they are perceived as a unit from the outside, and that they do something, i. e. they have the ability to *act*, and possibly to *react* to external stimuli and *interact* with the environment and other agents. The environment, which may include physical entities (infrastructures, geographical locations, etc.) and institutions (markets, regulatory systems, etc.), can also be modeled in terms of agents (e. g. a central bank, the order book of a stock exchange, etc.), whenever the conditions outlined above are met. When not, it should be thought of simply as a set of variables (say, temperature or business confidence).

The methodological issues are the real *litmus paper* of the competing approaches. According to one of the most quoted economic papers, Friedman [71], the ultimate goal of a positive science is to develop hypotheses that yield valid and meaningful *predictions* about actual phenomena. Not a word on predictions at the *meso*-level or on the realism of the hypotheses. Even the Occam rule is systematically ignored: e. g. to get a downward sloping aggregate demand curve, mainstream economics has to assume indifference curves which are: (i) defined only in the positive quadrant of commodity-bundle quantities; (ii) negatively sloped; (iii) complete; (iv) transitive, and (v) strictly convex, while ABM has to assume only the existence of reservation prices. Moreover, to properly aggregate from microbehavior, i. e. to get a well shaped aggregate demand from the individual ones, it has to be assumed that the propensity to consume out of income has to be homogeneous for all the agents (*homothetic* Engel curves) and that distribution is independent from relative prices. This methodology resembles the scientific procedure of the aruspexes, who predicted the future by reading the animals' bowels. The ABM methodology is bottom-up and focuses on the interaction between many heterogenous interacting agents, which might produce a statistical equilibrium, rather than a *natural* one as the mainstream approach assumes. The bottom-up approach models individual behavior according to simple behavioral rules; agents are allowed to have *local interaction* and to change the *individual rule* (*through adaptation*) as well as the *interaction nodes*. By aggregating, some *statistical regularity* emerges, which cannot be inferred from individual behavior *(self emerging regularities)*: this *emergent behavior* feeds back to the individual level *(downward causation)* thus establishing a macrofoundation of micro. As a consequence, each and every proposition may be falsified at *micro*, *meso* and *macro* levels. This approach opposes the axiomatic theory of economics, where the optimization procedure is the standard for a scientific, i. e. not ad-hoc, modeling procedure.

The agent-based methodology can also be viewed as a way to reconcile the two opposing philosophical perspectives of *methodological individualism* and *holism*. Having agents as the unit of analysis, ABM is deeply rooted in methodological individualism, a philosophical method aimed at explaining and understanding broad society-wide developments as the aggregation of decisions by individuals [13,172]. Methodological individualism suggests – in its most extreme (and erroneous) version – that a system can be understood by analyzing *separately* its constituents, the reductionist approach that the whole is nothing but the sum of its parts [51,127]. However, the ability to reduce everything to simple fundamental objects and laws does not imply the ability to start from those objects and laws and reconstruct the universe. In other terms, reductionism does not imply constructionism [2].

The Austrian school of economics championed the use of methodological individualism in economics in the twentieth century, of which Friederich von Hayek has been one of the main exponents. The legacy of Hayek to ABM and the complex system approach has been recognized [166]. Methodological individualism is also considered an essential part of modern neoclassical economics, with its analysis of collective action in terms of rational, utility-maximizing individuals: should the microfoundations in terms of individual rational behavior be abandoned, the Lucas Critique [114] would kick in. However, it is hard to recognize the imprinting of methodological individualism in the RA paradigm, which claims that the whole society can be analyzed in terms of the behavior of a single, representative, individual and forgets to apply to it the Lucas critique. On the other hand, focusing on aggregate phenomena arising from the bottom up [61] from the interaction of many different agents, ABM also adopts a holistic approach when it claims that these phenomena cannot be studied without looking at the entire context in which they are embedded. Indeed, holism is the idea that all the properties of a given system cannot be determined or explained by the sum of its component parts alone. Instead, the system as a whole determines in an important way that the parts behave. The general principle of holism was concisely summarized by Aristotle in his *Metaphysics*: "The whole is more than the sum of its parts", a *manifesto* of the complexity approach. However, ABM (and more in general complexity theory) should not be confused with general systems theory, an holistic approach developed in the 1950s and 1960s that in its most radical form argued that everything affects everything else: according to systems theory, phenomena that appear to have simple causes, such as unemployment, actually have a variety of complex causes – complex in the sense that the causes are interrelated, nonlinear, and difficult to determine [133]. Conversely, the complexity approach looks for *simple* rules that underpin complexity, an agenda that has been entirely transferred to ABM.

Also, ABM can be thought of as a bridge between methodological individualism and methodological holism. In agent-based models aggregate outcomes (the whole, e. g. the unemployment rate) are computed as the sum of individual characteristics (its parts, e. g. individual employment status). However, aggregate behavior can often be recognized as distinct from the behavior of the comprising agents, leading to the discovery of emergent properties. In this sense, the whole is more than – and different from – the sum of its parts. It might even be the case that the whole appears to act as if it followed a distinct logic, with its own goals and means, as in the example of a cartel of firms that act in order to influence the market price of a good. From the outside, the whole appears no different from a new agent type (e. g. a family, a firm). A new entity is born; the computational experiment has been successful in growing artificial societies from the bottom up [61].

This *bottom-up* approach to complexity consists in deducing the macroscopic objects (*macros*) and their phenomenological complex *ad-hoc* laws in terms of a multitude of elementary microscopic objects (*micros*) interacting by simple fundamental laws [154], and ABM provides a technique that allows one to systematically follow the birth of these complex macroscopic phenomenology. The *macros* at a specific scale can become the *micros* at the next scale.

Depending on the scope of the analysis, it is generally convenient to stop at some scale in the way down to reconstruct aggregate, top-level dynamics from the bottom up. When applied to economics, only a few levels (e. g. a *micro*, a *meso* and a *macro* level) are in general sufficient to provide a thorough understanding of the system. Defining the elementary units of analysis amounts to fixing the limits for the reductionist approach, which is not aprioristically discarded but rather integrated in the analysis. These units are in fact characterized by an inner structure that does not depend on the environment in which they are embedded. They can thus be analyzed separately.

The need for the ABM approach at any given scale is often linked to the existence of some underlying autocatalytic process at a lower level. *Autocatalytic processes* are dynamic processes with positive feedbacks, where the growth of some quantity is to some extent self-perpetuating, as in the case when it is proportional to its initial value. The importance of positive feedbacks has been recognized in the literature on increasing returns, in particular with respect to the possibility of multiple equilibria [151], since the time of Marshall. However, the traditional analysis is static, and does not address how an equilibrium out of several might be selected. Looking at the problem from a dynamic stochastic process perspective, selection is explained in terms of one set of small historical events magnified by increasing returns.

Moreover, the existence of an autocatalytic process implies that looking at the average, or most probable, behavior of the constituent units is non representative of the dynamics of the system: autocatalyticity insures that the behavior of the entire system is dominated by the elements with the highest auto-catalytic growth rate rather than by the typical or average element [154]. In presence of autocatalytic processes, even a small amount of individual heterogeneity invalidates any description of the behavior of the system in terms of its average element: the real world is controlled as much by the *tails* of distributions as by means or averages. We need to free ourselves from *average* thinking [3].

The fact that autocatalytic dynamics are scale invariant (i. e. after a transformation that multiplies all the variables by a common factor) is a key to understanding the emergence of scale invariant distributions of these variables (e. g. power laws), at an aggregate level. The relevance of scale free distributions in economics (e. g. of firm size, wealth, income, etc.) is now extensively recognized (Brock, 1999), and has been the subject of through investigation in the econophysics literature [120].

Additional Features of Agent-Based Models

We have so far introduced the three fundamental characteristics of ABM: there are agents that play the role of actors, there is no script or *Deus ex-machina* and the story is played live, i. e. it is computed. Following Epstein [58,59,60], we can further characterize the methodology, by enumerating a number of features that, although not necessary to define an agent-based model, are often present. These are:

Heterogeneity

While in analytical models there is a big advantage in reducing the ways in which individuals differ, the computational burden of ABM does not change at all if different values of the parameters (e. g. preferences, endowments, location, social contacts, abilities etc.) are specified for different individuals. Normally, a distribution for each relevant parameter is chosen, and this simply implies that a few parameters (those governing the distribution) are added to the model.

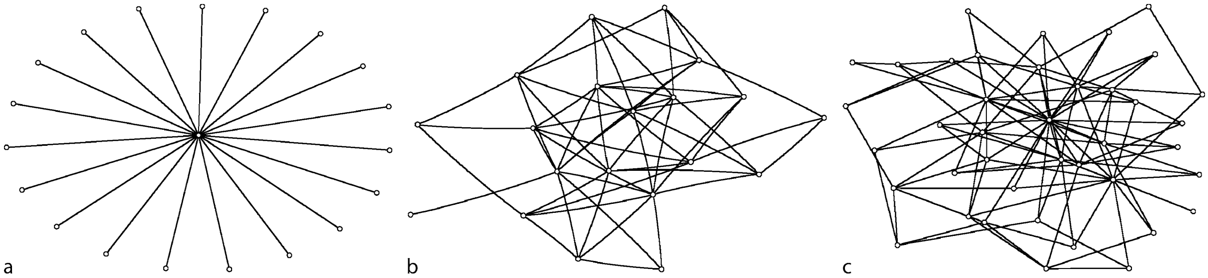
Explicit Space

This can be seen as specification of the previous point: individuals often differ in the physical place where they are located, and/or in the neighbors with whom they can or have to interact (which define the network structure of the model, see below).

Local Interaction

Again, this can be seen as a specification of the network structure connecting the agents. Analytical models often assume either global interaction (as in Walrasian markets), or very simple local interaction. ABM allow for much richer specifications. No direct interaction (only through prices) is allowed in the GE, while direct interaction (*local and stochastic*, usually [101]) is the rule for the complexity approach: figures 1a-c give a graphical representation of Walrasian, random and scale-free interaction respectively. Note that the empirical evidence supports the third case: hubs and power laws are the rule in the real world [38,52].

Actually, some neoclassical economists asked for an analysis of how social relations affect the allocation of resources (e. g., [12,107,134]). They went almost completely unheard, however, until the upsurge in the early 1990s of a brand new body of work aimed at understanding and modeling the social context of economic decisions, usually labeled *new social economics* or *social interaction economics* [56]. Models of social interactions (Manski [118] offers an operational classification of the channels through



Agent Based Models in Economics and Complexity, Figure 1
a a Walrasian GE representation; b a random graph; c a scale free graph in which several hubs can be identified

which the actions of one agent may affect those of other agents within a reference group) are generally able to produce several properties, such as *multiple equilibria* [34]; *non-ergodicity* and *phase transition* [54]; *equilibrium stratification* in social and/or spatial dimension [27,83]; the existence of a *social multiplier* of behaviors [84]. The key idea consists in recognizing that the social relationships in which individual economic agents are embedded can have a large impact on economic decisions. In fact, the social context impacts on individual economic decisions through several mechanisms. First, social norms, cultural processes and economic institutions may influence motivations, values, and tastes and, ultimately, make preferences endogenous [31]. Second, even if we admit that individuals are endowed with exogenously given preferences, the pervasiveness of information asymmetries in real-world economies implies that economic agents voluntarily share values, notions of acceptable behavior and socially based enforcement mechanisms in order to reduce uncertainty and favor coordination [50]. Third, the welfare of individuals may depend on some social characteristics like honor, popularity, stigma or status [41]. Finally, interactions not mediated by enforceable contracts may occur because of pure technological externalities in network industries [152] or indirect effects transmitted through prices (pecuniary externalities) in non-competitive markets [28], which may lead to coordination failures due to strategic complementarities [42].

Bounded Rationality

Interestingly, while in analytical models it is generally easier to implement some form of optimal behavior rather than solving models where individuals follow "reasonable" rules of thumb, or learn either by looking at what happened to others or what happened to them in the past, for ABM the opposite is true. However, it can be argued that *real* individuals also face the same difficulties in determining and following the optimal behavior, and are characterized by some sort of bounded rationality: "There are two components of this: bounded information and bounded computing power. Agents have neither global information nor infinite computational capacity. Although they are typically purposive, they are not global optimizers; they use simple rules based on local information" (p. 1588 in [59]).

The requirement on full rationality is indeed very strong, since it requires an infinite computational capacity (the ability of processing tons of data in a infinitesimal amount of time) and all the information. Moreover, according to the mainstream approach, information is complete and free for all the agents. Note that one of the assumptions in the Walrasian approach is that each agent has only private information: this is equivalent to say that *strategic behavior* about information collection and dissemination is ruled out and the collection of the whole set of the information is left to the market *via* the auctioneer (or a benevolent dictator [25]). Indeed, one could read the rational expectation "revolution" as the tentative to decentralize the price setting procedure by defenestrating the auctioneer. Limited information is taken into account, but the constraints have to affect every agent in the same way (the so-called Lucas' islands hypothesis) and the Greenwald–Stiglitz theorem [86] states that in this case the equilibrium is not even Pareto-constrained. If information is asymmetric or private, agents have to be heterogeneous and direct interaction has to be considered: this destroys the mainstream model and generates coordination failures.

On the contrary, agent-based models are build upon the hypothesis that agents have limited information. Once again, the ABM approach is much more parsimonious, since it only requires that the agents do not commit systematic errors. Moreover, given the limited information setting, the economic environment might change affecting, and being affected by, agents' behavior: individuals learn through experience and by interacting with other agents.

Non-equilibrium Dynamics

As we will explain in more details below, ABM are recursive models, in which the state of the system at time $t+1$ is computed starting from the state at time t. Hence, they allow the investigation of what happens all along the route, not only at the start and at the end of the journey. This point is, we believe, the most important. Brian Arthur (p. 1552 in [16]) offers an effective statement of its relevance for economic theory: "Standard neoclassical economics asks what agents' actions, strategies, or expectations are in equilibrium with (consistent with) the outcome or pattern these behaviors aggregatively create. Agent-based computational economics enables us to ask a wider question: how agents' actions, strategies or expectations might react to – might endogenously change with – the pattern they create. [...] This out-of-equilibrium approach is not a minor adjunct to standard economic theory; it is economics done in a more general way. [...] The static equilibrium approach suffers two characteristic indeterminacies: it cannot easily resolve among multiple equilibria; nor can it easily model individuals' choices of expectations. Both problems are ones of formation (of an equilibrium and of an 'ecology' of expectations, respectively), and when analyzed in formation – that is, out of equilibrium – these anomalies disappear".

As we have seen, continuous market clearing is *assumed* by the mainstream. It is a necessary condition to obtain "efficiency and optimality" and it is quite curious to read of a theory assuming the *explenandum*. In such a way, every out of equilibrium dynamics or path dependency is ruled out and initial conditions do not matter. The GE model assumes that transactions happen only after the vector of the equilibrium prices has been reached: instead of being the result of the exchange, it foresees it *par tatonnement* in a logical, fictitious time. Because the real markets operate in real, historical, time and the exchange process determines prices, the GE model is not able to describe any real economy [9]. Clower [39] suggested (resemblance Edgeworth, [57]) that exchange might happen out of equilibrium (*at false prices*). In such a case, agents will be quantity-rationed in their supply of-demand for: because of it, the intertemporal maximization problem has to be quantity-constraints (the so-called *Clower constraint*) and if the economy would reach equilibrium, it will be non-optimal and inefficient.

The requirement on rationality is also very strong, since it requires an infinite computational capacity (the ability of processing tons of data in a infinitesimal amount of time) and all the information. In fact, if information is limited, the outcome of a rational choice may be non-optimal. Once again, all the ABM approach is much more parsimonious, since it requires that the agents do not commit systematic errors. Moreover, given the limited information setting, the economic environment might change affecting, and being affected by, agents' behavior: learning and adaptive behavior are therefore contemplated.

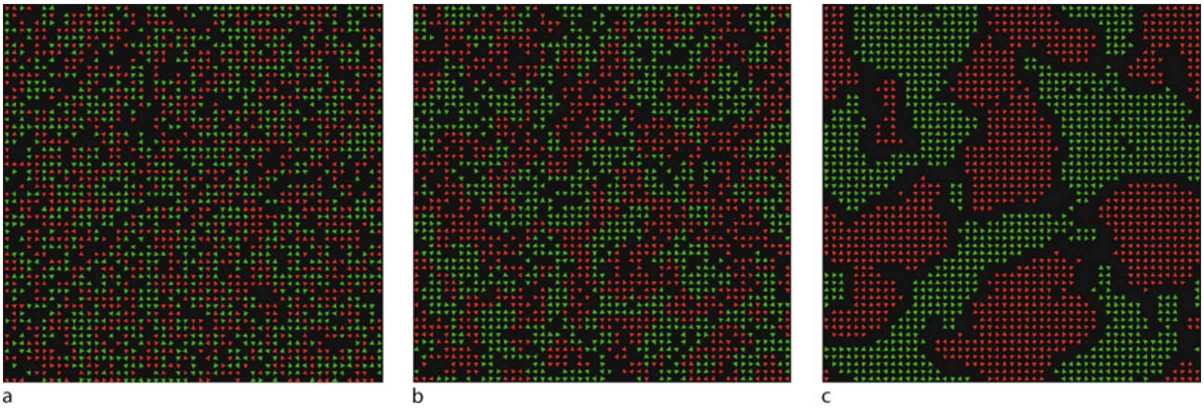
Finally, according to Beinhocker [26], the approaches differ also as regard *dynamics* (*Complex Systems* are open, dynamic, non-linear systems, far from equilibrium; *Mainstream* economics are closed, static, linear systems in equilibrium) and *evolution* (*Complex Systems* have an evolutionary process of differentiation, selection and amplification which provides the system with novelty and is responsible for its growth in order and complexity, while *Mainstream* has no mechanism for endogenously creating novelty, or growth in order and complexity.

An Ante Litteram Agent-Based Model: Thomas Schelling's Segregation Model

One of the early and most well known examples of an agent-based model is the segregation model proposed by Thomas Schelling [145,146], who in 2005 received the Nobel prize for his studies in game theory (surveys of more recent applications of ABM to economics can be found in [159,160,161,163]). To correctly assess the importance of the model, it must be evaluated against the social and historical background of the time. Up to the end of the 1960s racial segregation was institutionalized in the United States. Racial laws required that public schools, public places and public transportation, like trains and buses, had separate facilities for whites and blacks. Residential segregation was also prescribed in some States, although it is now widely recognized that it mainly came about through organized, mostly private efforts to ghettoize blacks in the early twentieth century – particularly the years between the world wars [63,126]. But if the social attitude was the strongest force in producing residential segregation, the Civil Rights movement of the 1960s greatly contributed to a change of climate, with the white population exhibiting increasing levels of tolerance. Eventually, the movement gained such strength to achieve its main objective, the abolition of the racial laws: this was sealed in the Civil Rights Act of 1968 which, among many other things, outlawed a wide range of discriminatory conduct in housing markets. Hence, both the general public attitude and the law changed dramatically during the 1960s. As a consequence, many observers predicted a rapid decline in housing segregation. The decline, however, was almost imperceptible. The question then was why this happened. Schelling's segregation model brought an answer, suggesting that small differences in tolerance level or initial location could trigger high level of segregation even without formal (i. e. legal) constraints, and even for decent levels of overall tolerance. In the model, whites and blacks are (randomly) located over a grid, each individual occupying one cell. As a consequence, each individual has at most eight neighbors (Moore neighborhood), located on adjacent cells. Preferences over residential patterns are represented as the maximum quota of racially different neighbors that an individual tolerates. For simplicity, we can assume that preferences are identical: a unique number defines the level of tolerance in the population. For example, if the tolerance level is 50% and an individual has only five neighbors, he would be satisfied if no more than two of his neighbors are racially different. If an individual is not satisfied by his current location, he tries to move to a different location where he is satisfied.

The mechanism that generates segregation is the following. Since individuals are initially located randomly on the grid, by chance there will be someone who is not satisfied. His decision to move creates two externalities: one in the location of origin and the other in the location of destination. For example, suppose a white individual decides to move because there are too many black people around. As he leaves, the ethnic composition of his neighborhood is affected (there is one white less). This increases the possibility that another white individual, who was previously satisfied, becomes eager to move. A similar situation occurs in the area of destination. The arrival of a white individual affects the ethnic composition of the neighborhood, possibly causing some black individual to become unsatisfied. Thus, a small non-homogeneity in the initial residential pattern triggers a chain effect that eventually leads to high levels of segregation. This mechanism is reinforced when preferences are not homogeneous in the population.

Figure 2, which shows the NETLOGO implementation of the Schelling model, exemplifies [173]. The left panel depicts the initial residential pattern, for a population of 2000 individuals, evenly divided between green and red, living on a 51 × 51 cells torus (hence the population density is 76.9%). Two values for the tolerance threshold are tested: in the first configuration, tolerance is extremely high (70%), while in the second it is significantly lower (30%), although at a level that would still be considered decent by many commentators. The initial residential pattern (obviously) shows no levels of segregation: every individual has on average 50% of neighbors of a different race. However, after just a few periods the equilibrium configurations of the middle (for a tolerance level of 70%) and right (for tolerance level of 30%) panels are obtained. The level of segregation is high: more than three quarters of neighbors are on average of the same racial group, even in case (b), when individuals are actually happy to live in a neighborhood dominated by a different racial group! Moreover, most people live in perfectly homogeneous clusters, with different ethnic clusters being often physically separated from each other by a no man's land. Only the relative mix brought by confining clusters keeps down the measure of overall segregation. Should the overall composition of the population be biased in favor of one



Agent Based Models in Economics and Complexity, Figure 2
NETLOGO implementation of Schelling's segregation model. a Initial (random) pattern. The average share of racially similar neighbors is roughly 50%. With a tolerance level of 70% (40%), less than 20% (more than 80%) of the individuals are not satisfied. b Final pattern. The average share of racially similar neighbors is 72.1%. Everyone is satisfied. c Final pattern. The average share of racially similar neighbors is 99.7%. Everyone is satisfied

ethnic group, we would clearly recognize the formation of ghettoes.

Note that the formation of racially homogeneous ethnic clusters and ghettoes is an emergent property of the system, which could hardly be deduced by looking at individual behavior alone, without considering the effects of interaction. Moreover, the clusters themselves could be considered as the elementary unit of analysis at a different, more aggregate level, and their behavior, e. g. whether they shrink, expand, merge or vanish, studied with respect to some exogenous changes in the environment. Not only a *property*, i. e. a statistical regularity, has emerged, but also a whole new *entity* can be recognized. However, this new entity is nothing else but a subjective interpretation by some external observer of an emergent property of the system.

The Development of Agent-Based Modeling

The early example of the segregation model notwithstanding, the development of agent-based computational economics is closely linked with the work conducted at the Santa Fe Institute for the study of complexity, a private, non-profit, independent research and education center founded in 1984 in Santa Fe, New Mexico. The purpose of the institute has been, since its foundation, to foster multidisciplinary collaboration in pursuit of understanding the common themes that arise in natural, artificial, and social systems. This unified view is the dominant theme of what has been called the *new science of complexity*.

The outcomes of this research program are well depicted in three books, all bearing the title *The economy as an evolving complex system* [4,17,29]. The following quotation, from the preface of the 1997 volume, summarizes very accurately the approach:

> "In September 1987 twenty people came together at the Santa Fe Institute to talk about 'the economy as a evolving, complex system'. Ten were theoretical economists, invited by Kenneth J. Arrow, and ten were physicists, biologists and computer scientists, invited by Philip W. Anderson. The meeting was motivated by the hope that new ideas bubbling in the natural sciences, loosely tied together under the rubric of 'the sciences of complexity', might stimulate new ways of thinking about economic problems. For ten days, economists and natural scientists took turns talking about their respective worlds and methodologies. While physicists grappled with general equilibrium analysis and non-cooperative game theory, economists tried to make sense of spin glass models, Boolean networks, and genetic algorithms. The meeting left two legacies. The first was the 1988 volume of essays; the other was the founding, in 1988, of the Economics Program at the Santa Fe Institute, the Institute's first resident research program. The Program's mission was to encourage the understanding of economic phenomena from a complexity perspective, which involved the development of theory as well as tools for modeling and for empirical analysis. [. . .] But just what is the complexity perspective in economics? That is not an easy question to answer. [. . .] Looking back over the developments in the past decade, and of the papers produced by the program, we believe that a coherent perspective – sometimes called the 'Santa Fe approach' – has emerged within economics."

The work carried out at the Santa Fe Institute greatly contributed to popularize the complexity approach to economics, although a similar line of research was initiated in Europe by chemists and physicists concerned with emergent structures and disequilibrium dynamics (more precisely, in Brussels by the group of the Nobel prize winner physical chemist Ilya Prigogine ([128]) and in Stuttgart by the group of the theoretical physicist Hermann Haken [91], as discussed in length by Rosser [141]).

Two main reasons can help explaining why the Santa Fe approach gained some visibility outside the restricted group of people interested in the complexity theory (perhaps contributing in this way to mount what Horgan [96,97], called an intellectual fad). Together, they offered an appealing suggestion of both what to do and how to do it. The first reason was the ability to present the complexity paradigm as a unitary perspective. This unitary vision stressed in particular the existence of feedbacks between *functionalities* and *objectives*: individual objectives determine to some extent the use and modification of existing functionalities, but functionalities direct to some extent the choice of individual objectives. It is this analytical focus that proved to be valuable in disciplines as diverse as the social sciences, the biological sciences and even architecture [70]. The second reason has to do with the creation of a specific simulation platform that allowed relatively inexperienced researchers to build their own toy models that, thanks to the enormous and sustained increase in commonly available computing power, could run quickly even on small PCs. This simulation platform was called SWARM [18], and consisted in a series of libraries that implemented many of the functionalities and technicalities needed to build an agent-based simulation, e. g. the schedule of the events, the passing of time and graphical widgets to monitor the simulation. In addition to offering a practical tool to write agent-based simulations, the SWARM approach proposed a protocol in simulation design, which the SWARM libraries exemplified.

The principles at the basis of the SWARM protocol are:

(i) The use of an *object-oriented programming* language (SWARM was first written in OBJECTIVE C, and later translated into JAVA), with different objects (and object types) being a natural counterpart for different agents (and agent types);

(ii) A *separate implementation* of the model and the tools used for monitoring and conducting experiments on the model (the so-called observer);

(iii) An architecture that allows nesting models one into another, in order to build a *hierarchy* of swarms – a swarm being a group of objects and a schedule of actions that the objects execute. One swarm can thus contain lower-level swarms whose schedules are integrated into the higher-level schedule.

A number of different simulation platforms that adhered to the SWARM protocol for simulation design have been proposed since, the most widespread being REPAST ([129]; see also [135]). However, other alternative approaches to writing agent-based models exist. Some rely on general-purpose mathematical software, like MATHEMATICA, MATLAB or MATCAD. Others, exemplified by the STARLOGO/NETLOGO experience [137], are based on the idea of an agent-based specific language.

Finally, despite the fact that ABM are most often computer models, and that the methodology could not develop in the absence of cheap and easy-to-handle personal computers, it is beneficial to remember that one of the most well-known agent-based models, the segregation model we have already described, abstracted altogether from the use of computers. As Schelling recalls, he had the original idea while seated on plane, and investigated it with paper and pencil. When he arrived home, he explained the rules of the game to his son and got him to move zincs and coppers from the child's own collection on a checkerboard, looking for the results: The dynamics were sufficiently intriguing to keep my twelve-year-old engaged p. 1643 in [148].

A Recursive System Representation of Agent-Based Models

Although the complexity theory is, above all, a mathematical concept, a rather common misunderstanding about agent-based simulations is that they are not as sound as mathematical models. In an often-quoted article, Thomas Ostrom [130] argued that *computer simulation* is a third symbol system in its own right, aside *verbal description* and *mathematics*: simulation is no mathematics at all (see [79]). An intermediate level of abstraction, according to this view, characterizes computer simulations: they are more abstract than verbal descriptions, but less abstract than pure mathematics. Ostrom (p. 384 in [130]) also argued that any theory that can be expressed in either of the first two symbol systems can also be expressed in the third symbol system. This implies that there might be verbal theories, which cannot be adequately expressed in the second symbol system of mathematics, but can be in the third [79].

This view has become increasingly popular among social simulators themselves, apparently because it offers a shield to the perplexity of the mathematicians, while hinting at a sort of superiority of computer simulations. Our opinion is that both statements are simply and plainly wrong. Agent-based modeling – and more in general simulation – *is* mathematics, as we argue in this paragraph. Moreover, the conjecture that any theory can be expressed via simulation is easily contradicted: think for instance of simulating Hegel's philosophical system.

Actually, agent-based simulations are nothing else but recursive systems [59,110], where the variables s that describe at time t the state of each individual unit are determined, possibly in a stochastic way, as a function of the past states s and some parameters a:

$$s_{i,t} = f_i(s_{i,t-1},\ s_{-i,t-1};\ a_i,\ a_{-i}\ ;\ t) \tag{1}$$

The individual state variables could include the memory of past values, as in the case when an unemployed person is characterized not only by the fact that he is unemployed, but also by when he last had a job. The function f_i and the parameters a_i determine individual behavior. They can possibly change over time, either in a random way or depending on some lagged variable or on higher-order parameters (as in the *Environment-Rule-Agent* framework of Gilbert and Terna [80]); when this is the case, their expression can simply be substituted for in Eq. (1). Equation (1) allows the recursive computation of the system: at any moment in time the state of each unit can be expressed as a (possibly stochastic) function of the initial values X_0 only, where X_0 includes the initial states and parameters of all the individual units:

$$s_{i,t} = g_i(X_0;\ t) \tag{2}$$

The aggregate state of the system is simply defined as

$$S_t = \sum_i s_{i,t} \tag{3}$$

Equilibrium in this system is described as a situation where the aggregate state S, or some other aggregate statistics Y computed on the individual states or the individual parameters are stationary.

Notice that this formalization describes both traditional dynamic micro models and agent-based simulations. In principle an agent-based model, not differently from traditional dynamic micro models, *could* be solved analytically. The problem is that the expressions involved quickly become unbearable, as (i) the level of heterogeneity, as measured by the distribution of the parameters a_i and functional forms f_i, increases; (ii) the amount of interaction, as measured by the dependency of $s_{i,t}$ on $s_{-i,t-1}$, increases; (iii) the functional forms f become more complicated, e. g. with the introduction of *if-else* conditions, etc.

Hence, the resort to numerical simulation. Traditional analytical models on the other hand must take great care that the system can be solved analytically, i. e. by symbolic manipulation. Hence the use of simple functions as the omnipresent Cobb–Douglas, the assumption of homogeneous units (that can then be replaced by a RA), the choice of simple interaction processes, often mediated by a centralized coordination mechanism. However, analytical tractability alone is a poor justification of any modeling choice. As the Nobel laureate Harry Markowitz wrote, "if we restrict ourselves to models which can be solved analytically, we will be modeling for our mutual entertainment, not to maximize explanatory or predictive power" (as reported in [112]). Restricting to analytically solvable *modls* – as they are called in the not sufficiently well-known paper by Axel Leijonhufvud [108] – looks dangerously close to the tale of the man who was searching for his keys under the light of a street lamp at night and, once asked if he had lost them there, he answered "No, but this is where the light is".

Analysis of Model Behavior

Being able to reach a close solution means that it is possible to connect inputs and outputs of the model, at any point in time, in a clear way: the *input-output transformation function*, or *reduced form*, implied by the *structural form* in which the model is expressed, is analytically obtained (e. g. the equilibrium expression of some aggregate variable of interest, as a function of the model parameters). Hence, theorems can be proved and laws expressed.

On the contrary, in a simulation model the reduced form remains unknown, and only *inductive* evidence about the input/output transformation implied by the model can be collected. Performing multiple runs of the simulation with different parameters does this. In other words, simulations suffer from the problem of stating general propositions about the dynamics of the model starting only from point observations. Since scientific explanations are generally defined as the derivation of general laws, which are able to replicate the phenomena of interests [93,94], simulations appear to be less scientific than analytical models. As Axelrod [19] points out, "like deduction, [a simulation] starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively". Induction comes at the moment of explaining the behavior of the model. It should be noted that although induction is used to obtain knowledge about the behavior of a given simulation model, the use of a simulation model to obtain knowledge about the behavior of the real world refers to the logical process of abduction [109,117]. Abduction [66,132], also called *inference to the best explanation*, is a method of reasoning in which one looks for the hypothesis that would best explain the relevant evidence, as in the case when the observation that the grass is wet allows one to suppose that it rained.

Being constrained to unveil the underlying input-output transformation function by repetitively sampling the parameter space, simulations cannot prove necessity, i. e. they cannot provide in the traditional sense necessary conditions for any behavior to hold. This is because nothing excludes a priori that the system will behave in a radically different way as soon as the value of some parameter is changed, while it is generally not possible to sample all values of the parameter space. In other words, the artificial data may not be representative of all outcomes the model can produce. While analytical results are conditional on the specific hypothesis made about the *model* only, simulation results are conditional both on the specific hypothesis of the model and the specific values of the *parameters* used in the simulation runs: each run of such a model yields is a sufficiency theorem, [yet] a single run does not provide any information on the robustness of such theorems [20].

The sampling problem becomes increasingly harder as the number of the parameters increase. This has been referred to as *the curse of dimensionality* [143]. To evaluate its implications, two arguments should be considered. The first one is theoretical: if the impossibility to gain a full knowledge of the system applies to the *artificial* world defined by the simulation model, it also applies to the *real* world. The real data generating process being itself unknown, stylized facts (against which all models are in general evaluated) could in principle turn wrong, at some point in time. From an epistemological point of view, our belief that the sun will rise tomorrow remains a probabilistic assessment. The second, and more decisive, consideration is empirical: we should not worry too much about the behavior of a model for particular evil combinations of the parameters, as long as these combinations remain extremely rare (one relevant exception is when rare events are the focus of the investigation, e. g. in risk management, see [150]). If the design of the experiments is sufficiently accurate, the problem of how imprecise is the estimated input-output transformation function becomes marginal:

> While the curse of dimensionality places a practical upper bound on the size of the parameter space that can be checked for robustness, it is also the case that vast performance increases in computer hardware are rapidly converting what was once perhaps a fatal difficulty into a manageable one [20].

In conclusion, extensive experimentation is the only way to get a full understanding of the simulation behavior. Sampling of the parameter space can be done either systematically, i. e. by grid exploration, or randomly. Following Leombruni et al. [111], we can further distinguish between two levels at which sampling can be done: a *global* level and a *local* level. Local sampling is conducted around some specific parameter configurations of interest, by letting each parameter vary and keeping all the others unchanged. This is known as *sensitivity analysis*, and is the equivalent to the study of the partial derivatives of the input-output transformation function in an analytical model.

As an example, Fig. 3 reports a plot of the equilibrium level of segregation in the Schelling model, for decreasing values of tolerance (left panel) and increasing population density (right panel). Tolerance level is sampled in the range $[0, .7]$ by increasing steps of .05, while population size is sampled in the range $[1000, 2000]$ by increasing steps of 100. To get rid of random effects (in the initial residential pattern and in the choice of a different location of unsatisfied individuals), 100 runs are performed for every value of the parameter being changed, and average outcomes are reported. This gives an idea of the local effects of the two parameters around the central parameter configuration where the population size is equal to 2000 and the tolerance level is equal to 70%.

For what concerns the effect of tolerance on segregation (left panel), it should be noted that the somewhat irregular shape of the relationship is a consequence not of the sample size but of the small neighborhood individuals take into consideration (a maximum of eight adjacent cells, as we have seen), and the discretization it brings. As the effect of population size on segregation (right panel) is concerned, it may seem at a first glance counter-intuitive that segregation initially diminishes, as the population density increases. This is due to the fact that clusters can separate more if there are more free locations. Of course, nothing excludes the possibility that these marginal effects are completely different around a different parameter configuration. To check whether this is the case, it is necessary either to repeat the sensitivity analysis around other configurations, or to adopt a multivariate perspective.

Allowing all parameters to change performs *global sampling*, thus removing the reference to any particular configuration. To interpret the results of such a global analysis, a relationship between inputs and outputs in the *artificial* data can be estimated, e. g.:

$$Y = m(X_0) \,. \tag{4}$$

Where Y is the statistics of interest (say, the Gini coefficient of wealth), computed in equilibrium, i. e. when it has reached stationarity, and X_0 contains the initial conditions and the structural parameters of the model: $X_0 = \{s_0, a\}$. If the (not necessary unique) steady state is independent of the initial conditions, Eq. 4 simplifies to:

$$Y = m(A) \,. \tag{5}$$

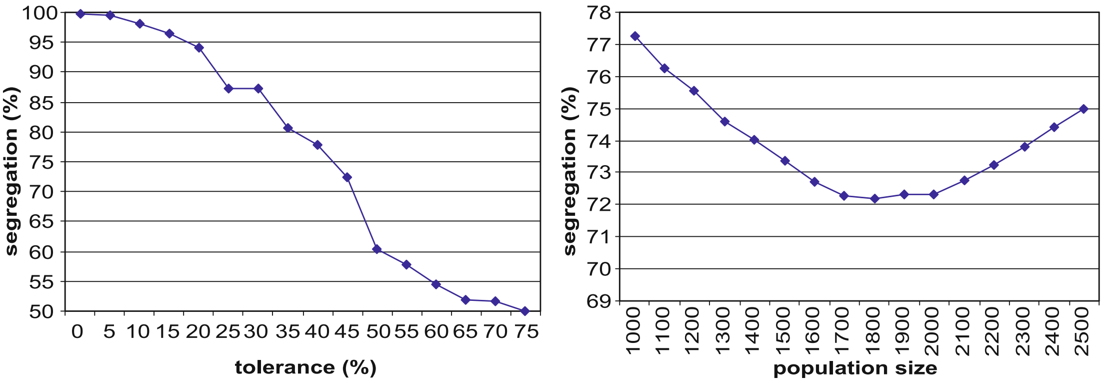
Where A contains only the parameters of the simulation. The choice of the functional form m to be estimated, which is sometimes referred to as *metamodel* [103] is to a certain extent arbitrary, and should be guided by the usual criteria for model specification for the analysis of real data.

As an example, we performed a multivariate analysis on the artificial data coming out of the Schelling's segregation model, by letting both the population size and the tolerance threshold to vary. Overall, 2115 parameter configurations are tested. After some data mining, our preferred specification is an OLS regression of the segregation level on a third order polynomial of the *tolerance* threshold, a second order polynomial of population *density*, plus an *interaction* term given by the product of tolerance and density. The interaction term, that turns out to be highly significant, implies that the local analysis of Fig. 3 has no general validity.

The regression outcome is reported in Table 1.

Such a model allows predicting the resulting segregation level for any value of the parameters. Of course, as the complexity of the model increases (e. g. leading to multiple equilibria) finding an appropriate meta-model becomes increasingly arduous.

Finally, let's remark that the curse of dimensionality strongly suggests that the flexibility in model specification characterizing agent-based models is to be used with care, never neglecting the KISS (*Keep it simple, Stupid*) prin-



Agent Based Models in Economics and Complexity, Figure 3
Sensitivity analysis for the Schelling's segregation model. Segregation is measured as the share of racially similar neighbors. The reference parameter configuration is population size = 2000, tolerance level = 40%

Agent Based Models in Economics and Complexity, Table 1
Regression results for Schelling's segregation model. Instead of repeating the experiment *n* times for each parameter configuration, in order to average out the random effects of the model, we preferred to test a number of different parameter configurations *n* times higher. Thus, population size is explored in the range [1000, 2000] by increasing steps of 10, and tolerance level is explored in the range [0, 7] by increasing steps of .05

Source	SS	df	MS
Model	666719.502	6	111119.917
Residual	12033.9282	2108	5.70869461
Total	678753.43	2114	321.075416

Number of obs = 2115
F(6, 2108) = 19465.03
Prob > F = 0.0000
R-squared = 0.9823
Adj R-squared = 0.9822
Root MSE = 2.3893

Segregation	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
tolerance	3.379668	.0819347	41.25	0.000	3.218987	3.54035
tolerance_2	−.0655574	.0013175	−49.76	0.000	−.0681411	−.0629737
tolerance_3	.0003292	6.73e−06	48.94	0.000	.000316	.0003424
density	−23.83033	3.274691	−7.28	0.000	−30.25229	−17.40837
density_2	20.05102	2.372174	8.45	0.000	15.39897	24.70306
interaction	−.1745321	.0153685	−11.36	0.000	−.2046712	−.144393
_cons	57.31189	1.957341	29.28	0.000	53.47336	61.15041

ciple. Schelling's segregation model is in this respect an example of simplicity, since it has but a few parameters: this is not incoherent with the complexity approach, since it stresses how simple behavioral rules can generate very complex dynamics.

Validation and Estimation

The previous section has dealt with the problem of interpreting the behavior of an agent-based model, and we have seen that this can be done by appropriately generating and analyzing *artificial* data. We now turn to the relationship between artificial and *real* data, that is (i) the problem of choosing the parameter values in order to have the behavior of the model being as close as possible to the real data, and (ii) the decision whether a model is good enough, which often entails a judgment on "how close" as close as possible is. The first issue is referred to as the problem of *calibration* or *estimation* of the model, while the second one is known as *validation.*

Note that all models have to be understood. Thus, for agent-based models analysis of the artificial data is always an issue. However, not all models have to be estimated or validated. Some models are built with a theoretical focus (e. g. Akerlof's market for lemons), and thus comparison with the real data is not an issue – although it could be argued that some sort of evaluation is still needed, although of a different kind.

Estimation

Although the terms calibration and estimation are sometimes given slightly different meanings (e. g. [105]), we agree with Hansen and Heckman (p. 91 in [92]) that "the distinction drawn between calibrating and estimating the parameters of a model is artificial at best. Moreover, the justification for what is called *calibration* is vague and confusing. In a profession that is already too segmented, the construction of such artificial distinctions is counterproductive."

Our understanding is that, too often, calibration simply refers to a sort of rough estimation, e. g. by means of visual comparison of the artificial and real data. However, not all parameters ought to be estimated by means of formal statistical methods. Some of them have very natural real counterparts and their value is known (e. g. the interest rate): the simulation is run with empirical data. Unknown parameters have on the other had to be properly estimated.

In analytical models the reduced form coefficients, e. g. the coefficients linking output variables to inputs, can be estimated in the real data. If the model is identified, there is a one-to-one relationship between the structural and the reduced form coefficients. Thus, estimates for the structural coefficients can be recovered. In a simulation model this can't be done. However, we could compare the outcome of the simulation with the real data, and change the structural coefficient values until the distance between the simulation output and the real data is minimized. This is called *indirect inference* [85], and is also applied to analytical models e. g. when it is not possible to write down the likelihood. There are many ways to compare real and artificial data. For instance, simple statistics can be computed both in real and in artificial data, and then aggregated in a unique measure of distance. Clearly, these statistics have to be computed just once in the real data (which does not change), and once every iteration until convergence in the artificial data, which depends on the value of the structural parameters. The change in the value of the parameters of each iteration is determined according to some optimization algorithm, with the aim to minimize the distance.

In the *method of simulated moments* different order of moments are used, and then weighted to take into account their uncertainty (while the uncertainty regarding the simulated moments can be reduced by increasing the number of simulation runs, the uncertainty in the estimation of the real, population moment on the basis of real sample data cannot be avoided). The intuition behind this is to allow parameters estimated with a higher degree of uncertainty to count less, in the final measure of distance between the real and artificial data [174]. Having different weights (or no weights at all) impinges on the efficiency of the estimates, not on their consistency. If the number of moments is equal to the number of structural parameters to be estimated, the model is just-identified and the minimized distance, for the estimated values of the parameters, is 0. If the number of moments is higher than the number of parameters the model is over-identified and the minimized distance is greater than 0. If it is lower it is under-identified. Another strategy is to estimate an *auxiliary model* both in the real and in the artificial data, and then compare the two sets of estimates obtained. The regression coefficients have the same role as the moments in the method of simulated moments: they are just a way of summarizing the data. Hence, if the number of coefficients in the auxiliary model is the same as the number of structural parameters to be estimated the model is just-identified and the minimized distance is 0. The specification of the auxiliary model is not too important. It can be proved that misspecification (e. g. omission of a relevant variable in the relationship to be estimated) only affects efficiency, while the estimates of the structural parameters remain consistent. A natural choice is of course the meta-model of Eq. 4.

Validation

A different issue is determining "how good" a model is. Of course, an answer to this question cannot be unique, but must be made in respect to some evaluation criterion. This in turn depends on the objectives of the analysis [62,102,111,164]. The need for evaluation of the model is no different in agent-based models and in traditional analytical models. However, like all simulations agent-based models require an additional layer of evaluation: the validity of the simulator (the program that simulates) relative to the model (*program validity*).

Assuming this is satisfied and the program has no bugs, Marks [121] formalizes the assessment of the *model validity* as follows: the model is said to be *useful* if it can exhibit at least some of the observed historical behaviors, *accurate* if it exhibits only behaviors that are compatible with those observed historically, and *complete* if it exhibits all the historically observed behaviors. In particular, letting **R** be the real world output, and **M** be the model output, four cases are possible:

a. No intersection between **R** and **M** ($\mathbf{R} \cap \mathbf{M} = \emptyset$): the model is *useless*;

b. **M** is a subset of **R** ($\mathbf{M} \subset \mathbf{R}$): the model is accurate, but *incomplete*;
c. **R** is a subset of **M** ($\mathbf{M} \supset \mathbf{R}$): the model is complete, but *inaccurate* (or *redundant*, since the model might tell something about what could yet happen in the world);
d. **M** is equivalent to **R** ($\mathbf{M} \Leftrightarrow \mathbf{R}$): the model is *complete* and *accurate*.

Of course, the selection of the relevant historical behaviors is crucial, and amounts to defining the criteria against which the model is to be evaluated. Moreover, the recognition itself of historical behavior passes through a process of analysis and simplification that leads to the identification of *stylized facts*, which are generally defined in stochastic terms. Thus, a model is eventually evaluated according to the extent to which it is able to statistically replicate the selected stylized facts.

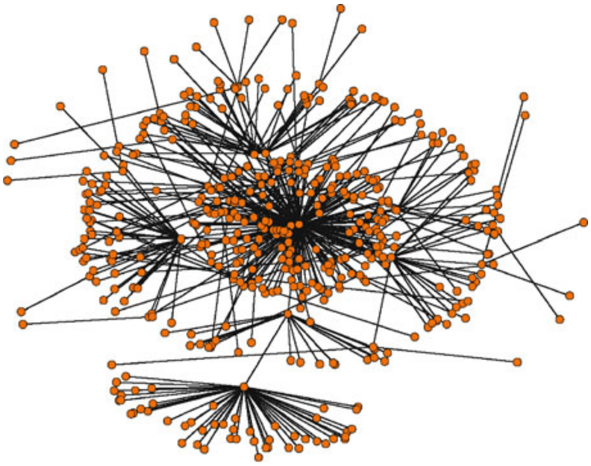
Finally, let's note that the behavior of the model might change significantly for different values of the parameters. Hence, the process of validation always regards both the *structure* of the model and the *values* of the parameters. This explains why and how validation and estimation are connected: as we have already noted, estimation is an attempt to make the behavior of the model as close as possible to real behavior; validation is a judgment on how far the two behaviors (still) are. A model where the parameters have not been properly estimated and are e. g. simple guesses can of course be validated. However, by definition its performance can only increase should the values of the parameters be replaced with their estimates.

The Role of Economic Policy

Before *economics* was *political economy*. According to the classical economists, the economic science has to be used to control the real economies and steer them towards desirable outcomes. If one considers the economic system as an analogue of the physical one, it is quite obvious to look for *natural* economic policy prescriptions (*one policy fits all*). This is the approach of mainstream (neoclassical) economists. There is a widespread opinion, well summarized by Brock and Colander [33], that, with respect to the economic policy analysis of the mainstream, (i) complexity does not add anything new to the box of tools. This point needs substantial corrections (see also the reflections by Durlauf [3]). The complexity approach showed us that the age of certainty ended with the non-equilibrium revolution, exemplified by the works of Prigogine. Considering the economy as an evolving (adaptive) system we have to admit that our understanding of it is limited (there is no room for *Laplace' demon* in complexity). Individual behavioral rules evolve according to their past performance: this provides a mechanism for an endogenous change of the environment. As a consequence the rational expectation hypothesis loses significance. However, agents are still rational in that they do what they can in order not to commit systematic errors [113]. In this setting there is still room for policy intervention outside the mainstream myth of a neutral and optimal policy. Because emergent facts are transient phenomena, policy recommendations are less certain, and they should be institution and historically oriented [65,170]. In particular, it has been emphasized that complex systems can either be extremely fragile and turbulent (a slight modification in some minor detail brings macroscopic changes), or relatively robust and stable: in such a context, policy prescriptions ought to be case sensitive.

In a heterogenous interacting agents environment, there is also room for an extension of the Lucas critique. It is well known that, according to it, because the underlying parameters are not policy-invariant any policy advice derived from large-scale econometric models that lack microfoundations would be misleading. The Lucas Critique implies that in order to predict the effect of a policy experiment, the so-called *deep parameters* (*preferences, technology* and *resource constraints*) that govern individual behavior have to be modeled. Only in this case it is possible to predict the behaviors of individuals, conditional on the change in policy, and aggregate them to calculate the macroeconomic outcome. But here is the trick: aggregation is a sum only if interaction is ignored. If non-price interactions (or other non-linearities) are important, then the interaction between agents may produce very different outcomes. Mainstream models focus on analytical solvable solutions: to get them, they have to simplify the assumptions e. g. using the RA approach or a Gaussian representation of heterogeneity. At the end, the main objective of these models is to fit the theory, not the empirics: how to explain, e. g., the scale-free network of the real economy represented in Fig. 1c by using the non interacting network of the mainstream model of Fig. 1a? At a minimum, one should recognize that the mainstream approach is a very primitive framework and, as a consequence, the economic policy recommendations derivable from it are very far from being adequate prescriptions for the real world.

Real economies are composed by millions of interacting agents, whose distribution is far from being stochastic or normal. As an example, Fig. 4 reports the distribution of the firms' trade-credit relations in the electronic-equipment sector in Japan in 2003 (see [47]). It is quite evident that there exist several hubs, i. e. firms with many connections: the distribution of the degree of connectivity



Agent Based Models in Economics and Complexity, Figure 4
Network of firms (electrical machinery and other machines sector, Japan). Source: De Masi et al. [47]

is *scale free*, i. e. there are a lot of firms with one or two links, and very a few firms with a lot of connections. Let us assume the Central Authority has to prevent a *financial collapse* of the system, or the spreading of a financial crisis (the so-called *domino effect*, see e. g. [104] and [157]). Rather than looking at the average risk of bankruptcy (in power law distributions the mean may even not exist, i. e. there is an empirical mean, but it is not stable), and to infer it is a measure of the stability of the system, by means of a network analysis the economy can be analyzed in terms of different interacting sub-systems, and local intervention can be recommended to prevent failures and their spread.

Instead of a helicopter drop of liquidity, one can make targeted interventions to a given agent or sector of activity: Fujiwara, [72], show how to calculate the probability of going bankrupt by *solo*, i. e. because of idiosyncratic elements, or *domino* effect, i. e. because of the failure or other agents with which there exist credit or commercial links.

One of the traditional fields of applications of economic policy is *redistribution*. It should be clear that a sound policy analysis requires a framework built without the RA straight jacket. A redistributive economic policy has to take into account that individuals are different: not only they behave differently, e. g. with respect to saving propensities, but they also have different fortunes: the so-called *St.Thomas* (13:12) effect (*to anyone who has, more will be given and he will grow rich; from anyone who has not, even what he has will be taken away*), which is the road to Paradise for Catholics, and to the power-law distribution of income and wealth for the econophysicists.

Gaffeo et al. [75], show that there is a robust link between firms' size distribution, their growth rate and GDP growth. This link determines the distributions of the amplitude frequency, size of recessions and expansion *etc.* Aggregate firms' size distribution can be well approximated by a *power law* [21,74], while sector distribution is still right skewed, but without scale-free characteristics [22]. Firms' growth rates are far from being *normal*: in the central part of the distribution they are tent shaped with very fat tails. Moreover, empirical evidence shows that exit is an inverse function of firms' age and size and proportional to financial fragility. In order to reduce the volatility of fluctuations, policy makers should act on the firms' size distribution, allowing for a growth of their capitalization, their financial solidity and wealth redistribution [48,49]. Since these emerging facts are policy sensitive, if the aggregate parameters change the shape of the curve will shift as well.

Differently from Keynesian economic policy, which theorizes aggregate economic policy tools, and mainstream neoclassical economics, which prescribes individual incentives because of the Lucas critique but ignores interaction which is a major but still neglected part of that critique, the ABM approach proposes a bottom up analysis. What generally comes out is not a one-size-fits-all policy since it depends on the general as well as the idiosyncratic economic conditions; moreover, it generally has to be conducted at different levels (from micro to meso to macro). In short, ABM can offer new answers to old unresolved questions, although it is still in a far too premature stage to offer definitive tools.

Future Directions

We have shown that mainstream approach to economics uses a methodology [71], which is so weak in its assumptions as to have been repeatedly ridiculed by the epistemologists [37], and dates back to the classical mechanical approach, according to which reductionism is possible. We have also seen that adopting the reductionist approach in economics is to say that agents *do not interact directly*: this is a very implausible assumption (billions of Robinson Crusoes who never meet Friday) and cannot explain the *emerging* characteristics of our societies, as witnessed by the empirical evidence. The reductionist approach of the mainstream is also theoretically incoherent, since it can be given no sound microfoundations [8,100].

In the fourth edition of his *Principles*, Marshall wrote, “The Mecca of the economist is biology”. What he meant to say was that, because economics deals with learning agents, evolution and change are the *granum salis* of our economic world. A theory built upon the issue of allocations of given quantities is not well equipped for the analysis of change. This allocation can be optimal only if there are no externalities (increasing returns, non-price interactions etc.) and information is complete, as in the case of the *invisible hand* parabola. In the history of science, there is a passage from a view emphasizing centralized intelligent design to a view emphasizing *self organized criticality* [27], according to which a system with many heterogenous interacting agents reaches a statistical aggregate equilibrium, characterized by the appearance of some (often scale free) stable distributions. These distributions are no longer optimal or efficient according to some welfare criterion: they are simply the natural outcome of individual interaction.

Because of the above-mentioned internal and external inconsistencies of the mainstream approach, a growing strand of economists is now following a different methodology based upon the analysis of systems with many heterogenous interacting agents. Their interaction leads to empirical regularities, which emerge from the system as a whole and cannot be identified by looking at any single agent in isolation: these emerging properties are, according to us, the main distinguishing feature of a complex system. The focus on interaction allows the scientist to abandon the heroic and unrealistic RA framework, in favor of the ABM approach, the *science of complexity* popularized by the SFI. Where did the Santa Fe approach go? Did it really bring a revolution in social science, as some of its initial proponents ambitiously believed? Almost twenty years and two "The economy as an evolving complex system" volumes later, Blume and Durlauf summarized this intellectual Odyssey as follows:

> "On some levels, there has been great success. Much of the original motivation for the Economics Program revolved around the belief that economic research could benefit from an injection of new mathematical models and new substantive perspectives on human behavior. [...] At the same time, [...] some of the early aspirations were not met" (Chaps. 1–2 in [29]).

It is probably premature to try to give definitive answers. For sure, ABM and the complexity approach are a very tough line of research whose empirical results are very promising (see e.g., Chaps. 2–3 in [77]). Modeling an agent-based economy however remains in itself a complex and complicated adventure.

Bibliography

1. Allen PM, Engelen G, Sanglier M (1986) Towards a general dynamic model of the spatial evolution of urban systems. In: Hutchinson B, Batty M (eds) Advances in urban systems modelling. North-Holland, Amsterdam, pp 199–220
2. Anderson PW (1972) More is different. Science 177:4047
3. Anderson PW (1997) Some thoughts about distribution in economics. In: Arthur WB, Durlaf SN, Lane D (eds) The economy as an evolving complex system II. Addison-Wesley, Reading
4. Anderson PW, Arrow K, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Redwood
5. Aoki M (1996) New approaches to macroeconomic modelling: evolutionary stochastic dynamics, multiple equilibria, and externalities as field effects. Cambridge University Press, Cambridge
6. Aoki M (2002) Modeling aggregate behaviour and fluctuations in economics. Cambridge University Press, Cambridge
7. Aoki M, Yoshikawa H (2006) Reconstructing macroeconomics. Cambridge University Press, Cambridge
8. Aoki M, Yoshikawa H (2007) Non-self-averaging in economic models. Economics Discussion Papers No. 2007-49, Kiel Institute for the World Economy
9. Arrow KJ (1959) Towards a theory of price adjustment. In: Abramovits M (ed) Allocation of economic resources. Stanford University Press, Stanford
10. Arrow KJ (1963) Social choice and individual values, 2nd edn. Yale University Press, New Haven
11. Arrow KJ (1964) The role of securities in the optimal allocation of risk-bearing. Rev Econ Stud 31:91–96
12. Arrow KJ (1971) A utilitarian approach to the concept of equality in public expenditures. Q J Econ 85(3):409–415
13. Arrow KJ (1994) Methodological individualism and social knowledge. Am Econ Rev 84:1–9
14. Arrow KJ, Debreu G (1954) Existence of an equilibrium for a competitive economy. Econometrica 22:265–290
15. Arthur WB (2000) Complexity and the economy. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Northampton
16. Arthur WB (2006) Out-of-equilibrium economics and agent-based modeling. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics, vol 2: Agent-Based Computational Economics, ch 32. North-Holland, Amsterdam, pp 1551–1564
17. Arthur WB, Durlauf S, Lane D (eds) (1997) The economy as an evolving complex system II. Addison-Wesley, Reading
18. Askenazi M, Burkhart R, Langton C, Minar N (1996) The swarm simulation system: A toolkit for building multi-agent simulations. Santa Fe Institute, Working Paper, no. 96-06-042
19. Axelrod R (1997) Advancing the art of simulation in the social sciences. In: Conte R, Hegselmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin, pp 21–40
20. Axtell RL (2000) Why agents? On the varied motivations for agent computing in the social sciences. Proceedings of the Workshop on Agent Simulation: Applications, Models and Tools. Argonne National Laboratory, Chicago
21. Axtell RL (2001) Zipf distribution of US firm sizes. Science 293:1818–1820
22. Axtell RL, Gallegati M, Palestrini A (2006) Common components in firms' growth and the scaling puzzle. Available at SSRN: http://ssrn.com/abstract=1016420
23. Bak P (1997) How nature works. The science of self-organized criticality. Oxford University Press, Oxford

24. Batten DF (2000) Discovering artificial economics. Westview Press, Boulder
25. Barone E (1908) Il ministro della produzione nello stato collettivista. G Econ 267–293, 391–414
26. Beinhocker ED (2006) The origin of wealth: Evolution, complexity, and the radical remaking of economics. Harvard Business School Press, Cambridge
27. Bénabou R (1996) Heterogeneity, stratification and growth: Macroeconomic implications of community structure and school finance. Am Econ Rev 86:584–609
28. Blanchard OJ, Kiyotaki N (1987) Monopolistic competition and the effects of aggregate demand. Am Econ Rew 77:647–666
29. Blume L, Durlauf S (eds) (2006) The economy as an evolving complex system, III. Current perspectives and future directions. Oxford University Press, Oxford
30. Blundell R, Stoker TM (2005) Heterogeneity and aggregation. J Econ Lit 43:347–391
31. Bowles S (1998) Endogenous preferences: The cultural consequences of markets and other economic institutions. J Econ Lit 36:75–111
32. Brock WA (1999) Scaling in economics: a reader's guide. Ind Corp Change 8(3):409–446
33. Brock WA, Colander D (2000) Complexity and policy. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Northampton
34. Brock WA, Durlauf SN (2001) Discrete choice with social interactions. Rev Econ Stud 68:235–260
35. Brock WA, Durlauf SN (2005) Social interactions and macroeconomics. UW-Madison, SSRI Working Papers n.5
36. Caballero RJ (1992) A Fallacy of composition. Am Econ Rev 82:1279–1292
37. Calafati AG (2007) Milton Friedman's epistemology UPM working paper n.270
38. Caldarelli G (2006) Scale-free networks. Complex webs in nature and technology. Oxford University Press, Oxford
39. Clower RW (1965) The keynesian counterrevolution: A theoretical appraisal. In: Hahn F, Brechling F (eds) The theory of interst rates. Macmillan, London
40. Cohen A, Harcourt G (2003) What ever happened to the Cambridge capital theory controversies. J Econ Perspect 17:199–214
41. Cole HL, Mailath GJ, Postlewaite A (1992) Social norms, savings behaviour, and growth. J Political Econ 100(6):1092–1125
42. Cooper RW (1999) Coordination games: Complementarities and macroeconomics. Cambridge University Press, Cambridge
43. Crutchfield J (1994) Is anything ever new? Considering emergence. In: Cowan G, Pines D, Meltzer D (eds) Complexity: Metaphors, models, and reality. Addison-Wesley, Reading, pp 515–537
44. Davis JB (2006) The turn in economics: Neoclassical dominance to mainstream pluralism? J Inst Econ 2(1):1–20
45. Debreu G (1959) The theory of value. Wiley, New York
46. Debreu G (1974) Excess demand functions. J Math Econ 1:15–23
47. De Masi G, Fujiwara Y, Gallegati M, Greenwald B, Stiglitz JE (2008) Empirical evidences of credit networks in Japan. mimeo
48. Delli Gatti D, Di Guilmi C, Gaffeo E, Gallegati M, Giulioni G, Palestrini A (2004) Business cycle fluctuations and firms' size distribution dynamics. Adv Complex Syst 7(2):1–18
49. Delli Gatti D, Di Guilmi C, Gaffeo E, Gallegati M, Giulioni G, Palestrini A (2005) A new approach to business fluctuations: Heterogeneous interacting agents, scaling laws and financial fragility. J Econ Behav Organ 56(4):489–512
50. Denzau AT, North DC (1994) Shared mental models: Ideologies and institutions. Kyklos 47(1):3–31
51. Descartes R (1637) Discours de la méthode pour bien conduire sa raison, et chercher la verité dans les sciences, tr. Discourse on Method and Meditations. The Liberal Arts Press, 1960, New York
52. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks from biological nets to the internet and the WWW. Oxford University Press, Oxford
53. Di Guilmi C, Gallegati M, Landini S (2007) Economic dynamics with financial fragility and mean-field interaction: a model. arXiv:0709.2083
54. Durlauf SN (1993) Nonergodic economic growth. Rev Econ Stud 60:349–366
55. Durlauf SN (1997) What should policymakers know about economic complexity? Wash Q 21(1):157–165
56. Durlauf SN, Young HP (2001) Social dynamics. The MIT Press, Cambridge
57. Edgeworth FY (1925) The pure theory of monopoly In: Papers relating to political economy. McMillan, London
58. Epstein JM (1999) Agent-based computational models and generative social science. Complexity 4:41–60
59. Epstein JM (2006) Remarks on the foundations of agent-based generative social science. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 34. North-Holland, Amsterdam, pp 1585–1604
60. Epstein JM (2006) Generative social science: Studies in agent-based computational modeling. Princeton University Press, New York
61. Epstein JM, Axtell RL (1996) Growing artificial societies: Social science from the bottom up. The MIT Press, Cambridge
62. Fagiolo G, Moneta A, Windrum P (2007) A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. Comput Econ 30:195–226
63. Farley R (1986) The residential segregation of blacks from whites: Trends, causes, and consequences. In: US Commission on Civil Rights, Issues in housing discrimination. US Commission on Civil Rights
64. Feller W (1957) An introduction to probability. Theory and its applications. Wiley, New York
65. Finch J, Orillard M (eds) (2005) Complexity and the economy: Implications for economy policy. Edward Elgar, Cheltenham
66. Flach PA, Kakas AC (eds) (2000) Abduction and induction. Essays on their relation and integration. Kluwer, Dordrecht
67. Flake GW (1998) The computational beauty of nature. The MIT Press, Cambridge
68. Foellmer H (1974) Random economies with many interacting agents. J Math Econ 1:51–62
69. Forni M, Lippi M (1997) Aggregation and the micro-foundations of microeconomics. Oxford University Press, Oxford
70. Frazer J (1995) An evolutionary architecture. Architectural Association Publications, London

71. Friedman M (1953) Essays in positive economics. University of Chicago Press, Chicago
72. Fujiwara Y (2006) Proceedings of the 9th Joint Conference on Information Sciences (JCIS), Advances in Intelligent Systems Research Series. Available at http://www.atlantis-press.com/publications/aisr/jcis-06/index_jcis
73. Gabaix X (2008) Power laws in Economics and Finance, 11 Sep 2008. Available at SSRN: http://ssrn.com/abstract=1257822
74. Gaffeo E, Gallegati M, Palestrini A (2003) On the size distribution of firms, additional evidence from the G7 countries. Phys A 324:117–123
75. Gaffeo E, Russo A, Catalano M, Gallegati M, Napoletano M (2007) Industrial dynamics, fiscal policy and R&D: Evidence from a computational experiment. J Econ Behav Organ 64:426–447
76. Gallegati M (1993) Composition effects and economic fluctuations. Econ Lett 44(1–2):123–126
77. Gallegati M, Delli Gatti D, Gaffeo E, Giulioni G, Palestrini A (2008) Emergent macroeconomics. Springer, Berlin
78. Gallegati M, Palestrini A, Delli Gatti D, Scalas E (2006) Aggregation of heterogeneous interacting agents: The variant representative agent framework. J Econ Interact Coord 1(1):5–19
79. Gilbert N (ed) (1999) Computer simulation in the social sciences, vol 42. Sage, Thousand Oaks
80. Gilbert N, Terna P (2000) How to build and use agent-based models in social science. Mind Soc 1:57–72
81. Gilbert N, Troitzsch K (2005) Simulation for the social scientist. Open University Press, Buckingham
82. Gintis H (2007) The dynamics of general equilibrium. Econ J 117:1280–1309
83. Glaeser E, Sacerdote B, Scheinkman J (1996) Crime and social interactions. Q J Econ 111:507–548
84. Glaeser J, Dixit J, Green DP (2002) Studying hate crime with the internet: What makes racists advocate racial violence? J Soc Issues 58(122):177–194
85. Gourieroux C, Monfort A (1997) Simulation-based econometric methods. Oxford University Press, Oxford
86. Greenwald B, Stiglitz JE (1986) Externalities in economies with imperfect information and incomplete markets. Q J Econ 101(2):229–264
87. Grossman SJ, Stiglitz JE (1976) Information and competitive price systems. Am Econ Rev 66:246–253
88. Grossman SJ, Stiglitz JE (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70(3):393–408
89. Guesnerie R (1993) Successes and failures in coordinating expectations. Eur Econ Rev 37:243–268
90. Hahn F (1982) Money and inflation. Blackwell, Oxford
91. Haken H (1983) Synergetics. Nonequilibrium phase transitions and social measurement, 3rd edn. Springer, Berlin
92. Hansen L, Heckman J (1996) The empirical foundations of calibration. J Econ Perspect 10:87–104
93. Hempel CV (1965) Aspects of scientific explanation. Free Press, London
94. Hempel CV, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175
95. Hildenbrand W, Kirman AP (1988) Equilibrium analysis: Variations on the themes by edgeworth and walras. North-Holland, Amsterdam
96. Horgan J (1995) From complexity to perplexity. Sci Am 272:104
97. Horgan J (1997) The end of science: Facing the limits of knowledge in the twilight of the scientific age. Broadway Books, New York
98. Jerison M (1984) Aggregation and pairwise aggregation of demand when the distribution of income is fixed. J Econ Theory 33(1):1–31
99. Kirman AP (1992) Whom or what does the representative individual represent. J Econ Perspect 6:117–136
100. Kirman AP (1996) Microfoundations – built on sand? A review of Maarten Janssen's microfoundations: A Critical Inquiry. J Econ Methodol 3(2):322–333
101. Kirman AP (2000) Interaction and markets. In: Gallegati M, Kirman AP (eds) Beyond the representative agent. Edward Elgar, Cheltenham
102. Kleijnen JPC (1998) Experimental design for sensitivity analysis, optimization, and validation of simulation models. In: Banks J (ed) Handbook of simulation. Wiley, New York, pp 173–223
103. Kleijnen JPC, Sargent RG (2000) A methodology for the fitting and validation of metamodels in simulation. Eur J Oper Res 120(1):14–29
104. Krugman P (1998) Bubble, boom, crash: theoretical notes on Asia's crisis. mimeo
105. Kydland FE, Prescott EC (1996) The computational experiment: An econometric tool. J Econ Perspect 10:69–85
106. Lavoie D (1989) Economic chaos or spontaneous order? Implications for political economy of the new view of science. Cato J 8:613–635
107. Leibenstein H (1950) Bandwagon, snob, and veblen effects in the theory of consumers' demand. Q J Econ 64:183–207
108. Leijonhufvud A (1973) Life among the econ. Econ Inq 11:327–337
109. Leombruni R (2002) The methodological status of agent-based simulations, LABORatorio Revelli. Working Paper No. 19
110. Leombruni R, Richiardi MG (2005) Why are economists sceptical about agent-based simulations? Phys A 355:103–109
111. Leombruni R, Richiardi MG, Saam NJ, Sonnessa M (2005) A common protocol for agent-based social simulation. J Artif Soc Simul 9:1
112. Levy M, Levy H, Solomon S (2000) Microscopic simulation of financial markets. In: From Investor Behavior to Market Phenomena. Academica Press, New York
113. Lewontin C, Levins R (2008) Biology under the influence: Dialectical essays on the coevolution of nature and society. Monthly Review Press, US
114. Lucas RE (1976) Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series, vol 1, pp 19–46
115. Lucas RE (1987) Models of business cycles. Blackwell, New York
116. Lucas RE, Sargent T (1979) After keynesian macroeconomics. Fed Reserv Bank Minneap Q Rev 3(2):270–294
117. Magnani L, Belli E (2006) Agent-based abduction: Being rational through fallacies. In: Magnani L (ed) Model-based reasoning in science and engineering, Cognitive Science, Epistemology, Logic. College Publications, London, pp 415–439
118. Manski CF (2000) Economic analysis of social interactions. J Econ Perspect 14:115–136
119. Mantel R (1974) On the characterization of aggregate excess demand. J Econ Theory 7:348–353

120. Mantegna RN, Stanley HE (2000) An introduction to econophysics. Cambridge University Press, Cambridge
121. Marks RE (2007) Validating Simulation Models: A general framework and four applied examples. Comput Econ 30:265–290
122. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
123. Mas-Colell A, Whinston MD, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
124. Miller JH, Page SE (2006) Complex adaptive systems: An introduction to computational models of social life. Princeton University Press, New York
125. Mirowski P (1989) More heat than light. Cambridge University Press, Cambridge
126. Muth RF (1986) The causes of housing segregation. US Commission on Civil Rights, Issues in Housing Discrimination. US Commission on Civil Rights
127. Nagel E (1961) The structure of science. Routledge and Paul Kegan, London
128. Nicolis G, Prigogine I (1989) Exploring complexity: An introduction. WH Freeman, New York
129. North MJ, Howe TR, Collier NT, Vos JR (2005) Repast simphony runtime system. In: Macal CM, North MJ, Sallach D (eds) Proceedings of the agent 2005 Conference on Generative Social Processes, Models, and Mechanisms, 13–15 Oct 2005
130. Ostrom T (1988) Computer simulation: the third symbol system. J Exp Soc Psycholog 24:381–392
131. Page S (1999) Computational models from A to Z. Complexity 5:35–41
132. Peirce CS (1955) Abduction and induction. In: J Buchler (ed) Philosophical writings of peirce. Dover, New York
133. Phelan S (2001) What is complexity science, really? Emergence 3:120–136
134. Pollack R (1975) Interdependent preferences. Am Econ Rev 66:309–320
135. Railsback SF, Lytinen SL, Jackson SK (2006) Agent-based simulation platforms: Review and development recommendations. Simulation 82:609–623
136. Rappaport S (1996) Abstraction and unrealistic assumptions in economics. J Econ Methodol 3(2):215–36
137. Resnick M (1994) Turtles, termites and traffic jams: Explorations in massively parallel microworlds. MIT, Cambidge
138. Richter MK, Wong K (1999) Non-computability of competitive equilibrium. Econ Theory 14:1–28
139. Rioss Rull V (1995) Models with heterogeneous agents. In: Cooley TF (ed) Frontiers of business cycle research. Princeton University Press, New York
140. Rosser JB (1999) On the complexities of complex economic dynamics. J Econ Perspect 13:169–192
141. Rosser JB (2000) Integrating the complexity vision into the teaching of mathematical economics. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Cheltenham, pp 209–230
142. Rosser JB (2003) Complexity in economics. Edward Elgar, Cheltenham
143. Rust J (1997) Using randomization to break the curse of dimensionality. Econometrica 65:487–516
144. Saari DG (1995) Mathematical complexity of simple economics. Notices Am Math Soc 42:222–230
145. Schelling TC (1969) Models of segregation. Am Econ Rev 59:488–493
146. Schelling TC (1971) Dynamic models of segregration. J Math Sociol 1:143–186
147. Schelling TC (1978) Micromotives and macrobehaviour. W.W. Norton, New York
148. Schelling TC (2006) Some fun, thirty-five years ago. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 37. North-Holland, Amsterdam, pp 1639–1644
149. Schumpeter JA (1960) History of economic analysis. Oxford University Press, Oxford
150. Segre-Tossani L, Smith LM (2003) Advanced modeling, visualization, and data mining techniques for a new risk landscape. Casualty Actuarial Society, Arlington, pp 83–97
151. Semmler W (2005) Introduction (multiple equilibria). J Econ Behav Organ 57:381–389
152. Shy O (2001) The economics of network industries. Cambridge University Press, Cambridge
153. Smith A (1776/1937) The wealth of nations. Random House, New York
154. Solomon S (2007) Complexity roadmap. Institute for Scientific Interchange, Torino
155. Sonnenschein H (1972) Market excess demand functions. Econometrica 40:549–563
156. Stiglitz JE (1992) Methodological issues and the new keynesian economics. In: Vercelli A, Dimitri N (eds) Macroeconomics: A survey of research strategies. Oxford University Press, Oxford, pp 38–86
157. Stiglitz JE (2002) Globalization and its discontents. Northon, New York
158. Stoker T (1995) Empirical approaches to the problem of aggregation over individuals. J Econ Lit 31:1827–1874
159. Tesfatsion L (ed) (2001) Special issue on agent-based computational economics. J Econ Dyn Control 25
160. Tesfatsion L (ed) (2001) Special issue on agent-based computational economics. Comput Econ 18
161. Tesfatsion L (2001) Agent-based computational economics: A brief guide to the literature. In: Michie J (ed) Reader's guide to the social sciences. Fitzroy-Dearborn, London
162. Tesfatsion L (2002) Agent-based computational economics: Growing economies from the bottom up. Artif Life 8:55–82
163. Tesfatsion L (2006) Agent-based computational economics: A constructive approach to economic theory. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 16. North-Holland, Amsterdam, pp 831–880
164. Troitzsch KG (2004) Validating simulation models. In: Horton G (ed) Proceedings of the 18th european simulation multiconference. Networked simulations and simulation networks. SCS Publishing House, Erlangen, pp 265–270
165. Vriend NJ (1994) A new perspective on decentralized trade. Econ Appl 46(4):5–22
166. Vriend NJ (2002) Was Hayek an ace? South Econ J 68:811–840
167. Velupillai KV (2000) Computable economics. Oxford University Press, Oxford
168. Velupillai KV (2002) Effectivity and constructivity in economic theory. J Econ Behav Organ 49:307–325
169. Velupillai KV (2005) The foundations of computable general equilibrium theory. In: Department of Economics Working Papers No 13. University of Trento

170. Velupillai KV (2007) The impossibility of an effective theory of policy in a complex economy. In: Salzano M, Colander D (eds) Complexity hints for economic policy. Springer, Milan
171. von Hayek FA (1948) Individualism and economic order. University of Chicago Press, Chicago
172. von Mises L (1949) Human action: A treatise on economics. Yale University Press, Yale
173. Wilensky U (1998) NetLogo segregation model. Center for connected learning and computer-based modeling. Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/models/Segregation
174. Winker P, Gilli M, Jeleskovic V (2007) An objective function for simulation based inference on exchange rate data. J Econ Interact Coord 2:125–145
175. Wooldridge M (2001) An introduction to multiagent systems. Wiley, New York

Bayesian Methods in Non-linear Time Series

OLEG KORENOK
Department of Economics, VCU School of Business,
Richmond, USA

Article Outline

Glossary

Autoregressive model describes a stochastic process as a weighted average of its previous values and a stochastic error term.

Threshold autoregressive model is an autoregressive model in which parameters change depending on the time index or the previous values of the process.

Markov-switching autoregressive model is an autoregressive model in which parameters change over time depending on an unobserved Markov chain.

Prior distribution summarizes the information about the parameters of interest after observing the data.

Posterior distribution summarizes the information about the parameters of interest after observing the data.

Definition of the Subject

Economic fluctuations display definite nonlinear features. Recessions, wars, financial panics, and varying government policies change the dynamics of almost all macroeconomic and financial time series. In the time series literature, such events are modeled by modifying the standard linear autoregressive (abbreviated, AR) model

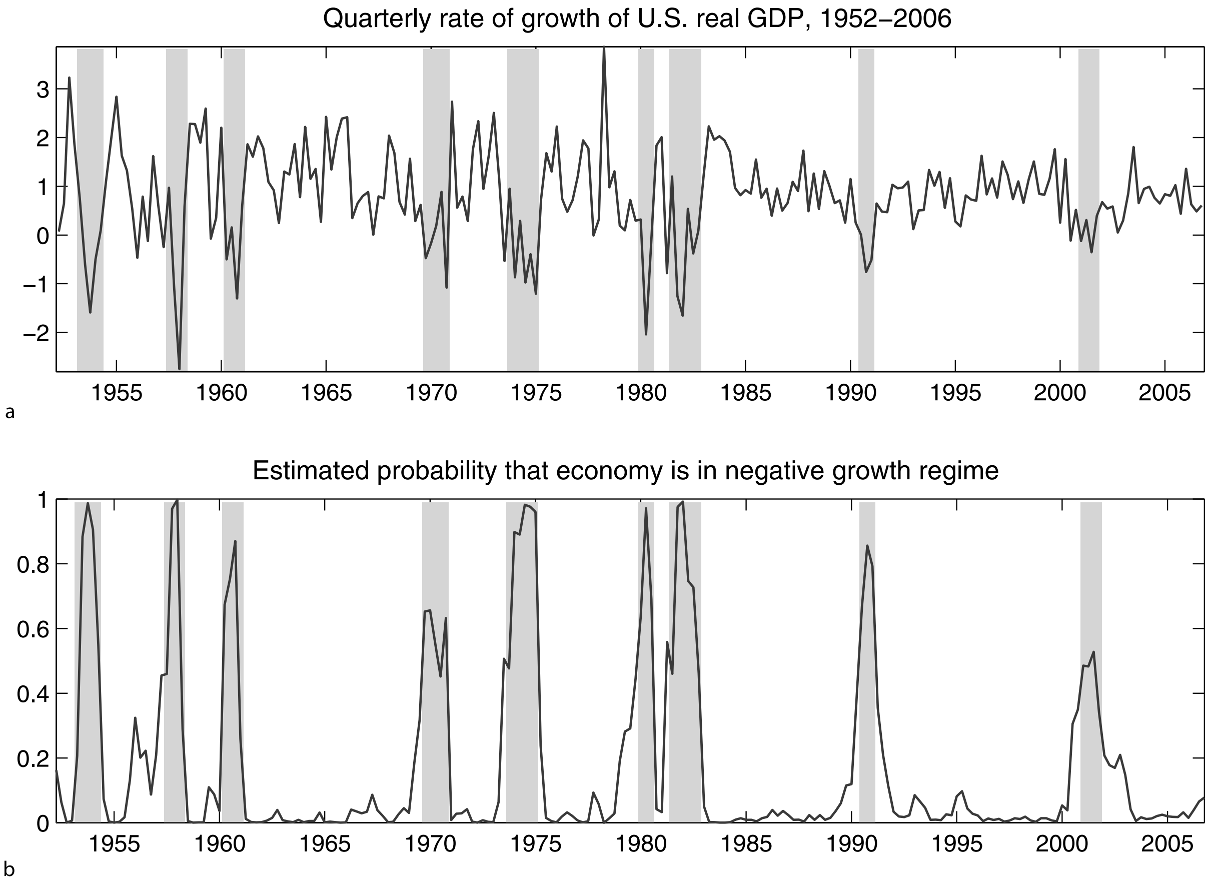
$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \,,$$

where y_t is a covariance stationary process, ϵ_t is an independent and identically distributed noise process, $\epsilon_t \sim i.i.d. N(0, \sigma^2)$, and the parameters c, ϕ_i, and σ^2 are fixed over time. In particular, the literature assumes that y_t follows two or more regimes. The three most commonly used nonlinear models differ in their description of the transition between regimes. In the threshold autoregressive (abbreviated, TAR) model, regime changes abruptly; in the smooth threshold autoregressive (abbreviated, STAR) model, regime changes slowly. Nevertheless, in both models the regime change depends on the time index or lagged values of y_t. In the Markov-switching autoregressive (abbreviated, MAR) model, however, the regime change depends on the past values of an unobserved random variable, the state of the Markov chain, and possibly the lagged values of y_t.

Arguably, the best-known example of the nonlinear time series model is the model of cyclical fluctuations of the US economy. It was first introduced and estimated by Hamilton [45] for quarterly US real Gross National Product over the 1952(II)–1984(IV) period. The model has two discrete regimes. The first regime is associated with a positive 1.2% growth rate and the second regime is associated with a negative −0.4% growth rate. Against his original motivation to find decade-long changes in growth rate trends for the US economy, Hamilton finds that negative growth regimes occur at the business cycle frequency. Positive growth regimes last, on average, 10 quarters, and negative growth regimes last, on average, 4 quarters. Moreover, he finds that the estimated regimes coincide closely with the official National Bureau of Economic Research (abbreviated, NBER) recession dates.

Figure 1 illustrates Hamilton's results for the extended 1952(II)–2006(IV) sample. Panel (a) shows the quarterly growth rate of the US real Gross Domestic Product, currently the more common measure of output; panel (b) plots the estimated probability that the US economy is in a negative growth regime. The shaded regions represent recessionary periods as determined informally and with some delay by the NBER: It took nine months for the NBER's Business Cycle Dating Committee to determine the latest peak of the US economy, which occurred in March 2001 but was officially announced in November 2001. Even though the NBER dates were not used in the model, the periods with high probability of a negative growth rate coincide almost perfectly with the NBER dates.

In addition to the formal recession dating methodology, Hamilton [45] presents clear statistical evidence for the proposition that the US business cycle is asymmetric: Behavior of output during normal times, when labor, capital, and technology determine long-run economic growth, is distinct from behavior during recessions, when all these factors are underutilized.



Bayesian Methods in Non-linear Time Series, Figure 1
Output growth and recession probabilities

Introduction

Hamilton's paper triggered an explosion of interest in nonlinear time series. The purpose of this paper is to give a survey of the main developments from the Bayesian perspective. The Bayesian framework treats model parameters as random variables and interprets probability as a degree of belief about particular realizations of a random variable conditional on available information. Given the observed sample, the inference updates prior beliefs, formulated before observing the sample, into posterior beliefs using Bayes' theorem

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} ,$$

where y is the sample observations $y = (y_1, \dots, y_T)$, θ is the vector of parameters $\theta = (c, \phi_1, \dots, \phi_p, \sigma^2)$, $\pi(\theta)$ is the prior distribution that describes beliefs prior to observing the data, $f(y|\theta)$ is the distribution of the sample conditional on the parameters, $f(y)$ is the marginal distribution of the sample, and $p(\theta|y)$ is the posterior distribution that describes the beliefs after observing the sample. Zellner [100], Bauwens, Lubrano, and Richard [7], Koop [58], Lancaster [61], and Geweke [39] cover Bayesian econometrics extensively and provide excellent introductions to relevant computational techniques.

We review the three most commonly used nonlinear models in three separate sections. We start each section by describing a baseline model and discussing possible extensions and applications (Matlab implementation of baseline models is available at http://www.people.vcu.edu/~okorenok/share/mlab.zip). Then we review the choice of prior, inference, tests against the linear hypothesis, and conclude with models selection. A short discussion of recent progress in incorporating regime changes into theoretical macroeconomic models concludes our survey.

Our survey builds on reviews of the TAR and STAR models in Tong [95], Granger and Terasvirta [41], Terasvirta [90], Bauwens, Lubrano, and Richard [7], Lubrano [63], Potter [74], Franses and van Dijk [34], van Dijk, Terasvirta, and Franses [98], and on reviews of the MAR models in Hamilton [46], Potter [74], and Kim and Nelson [51].

We limit our survey of nonlinear models only to the TAR, STAR, and MAR models. For a reader interested in a wider range of time series models from a Bayesian prospective, we recommend Steel's [84] survey: He overviews linear, as well as nonlinear, and parametric, as well as nonparametric, models.

Threshold Autoregressive Model

A threshold regression was introduced by Quandt [75] and was extended to the threshold autoregressive model by Tong [92,93] and Tong and Lim [94]. Tong [95] had a great impact on popularizing TAR models.

We limit our baseline model to a single switching variable z_t. The choice of the switching variable depends on the purpose of the investigation. For the analysis of structural breaks at an unknown point in time, Perron and Vogelsang [70], as well as DeJong [24], among many others, use the time index ($z_t = t$). For the purpose of prediction, Geweke and Terui [37], Chen and Lee [15], and others, use a lagged value of the time series ($z_t = y_{t-d}$), the self-exciting threshold autoregressive (abbreviated, SETAR) model.

In our discussion, the number of lags in the model p and a delay d is fixed. We also limit the baseline model to the homoscedastic case so that the variance of ϵ_t is constant in both regimes.

Introducing a more general notation, $x'_t = (1, y_{t-1}, \dots, y_{t-p})$, $\beta' = (c, \phi_1, \dots, \phi_p)$, the two-regime TAR model becomes

$$\begin{aligned} y_t &= x'_t\beta_1 + \epsilon_t \quad \text{if} \quad z_t < \tau \quad \text{(first regime)},\\ y_t &= x'_t\beta_2 + \epsilon_t \quad \text{if} \quad z_t \geq \tau \quad \text{(second regime)}\,, \end{aligned}$$

or more succinctly

$$y_t = [1 - I_{[\tau,\infty)}(z_t)]x'_t\beta_1 + I_{[\tau,\infty)}(z_t)x'_t\beta_2 + \epsilon_t\,, \quad (1)$$

where $I_A(x)$ is an indicator function that is equal to one if $x \in A$, in particular $I_{[\tau,\infty)}(z_t) = 1$ if $z_t \in [\tau, \infty)$. The indicator function introduces the abrupt transition between regimes. It is convenient to rewrite the model in a more compact form

$$y_t = x'_t(\tau)\beta + \epsilon_t\,, \quad (2)$$

where $x'_t(\tau) = (x'_t, I_{[\tau,\infty)}(z_t)x'_t)$ and $\beta' = (\beta'_1, \delta')$ with $\delta = \beta_2 - \beta_1$.

If the number of observations in regime i is less than or equal to the number of parameters, we cannot estimate parameters, or the model is not identified. In the Bayesian inference, we resolve the identification problem by restricting the region of possible parameter values to the one where the number of observations per regime is greater than the number of regressors.

The baseline model can be extended in several ways. First, we can allow the variance of the error term to differ in each regime. In this case, we rescale the data and introduce an additional parameter $\phi = \sigma_2^2/\sigma_1^2$, as in Lubrano [63]. Second, we can allow the number of lags to differ in each regime. Then p equals to $\max\{p_1, p_2\}$.

A more substantial change is required if we want to increase the number of regimes r. We can either use a single transition variable

$$y_t = x_t\beta_{i}(t) + \sigma_i(t)\epsilon_t\,,$$

where $i(t) = 1$ if $z_t < \tau_1$, $i(t) = 2$ if $\tau_1 \leq z_t < \tau_2$, $\dots$, $i(t) = r$ if $\tau_{r-1} \leq z_t$; or we can use a combination of two (or more) transition variables as in Astatkie, Watts, and Watt [5], where first stage transition is nested in the second stage transition

$$\begin{aligned} y_t = {} & [(1 - I_{[\tau_1,\infty)}(z_{1t}))x'_t\beta_1 + I_{[\tau_1,\infty)}(z_{1t})x'_t\beta_2]\\ & \cdot [1 - I_{[\tau_2,\infty)}(z_{2t})]\\ & + [(1 - I_{[\tau_1,\infty)}(z_{1t}))x'_t\beta_3 + I_{[\tau_1,\infty)}(z_{1t})x'_t\beta_4]\\ & \cdot I_{[\tau_2,\infty)}(z_{2t}) + \epsilon_t\,, \end{aligned}$$

nested TAR model.

Also, we can treat either the choice of number of lags, the delay, or the number of regimes as an inference problem. Then p, d, and r are added to the vector of the model parameters, as in Geweke and Terui [37] and Koop and Potter [57].

Finally, the univariate TAR model can be extended to describe a vector of time series as in Tsay [96]. The n dimensional two-regime TAR model can be specified in a manner similar to Eq. (1) as

$$\begin{aligned} Y_t = {} & [1 - I_{[\tau,\infty)}(z_t)](C_1 + \Phi_{11}Y_{t-1} + \cdots + \Phi_{1p}Y_{t-p})\\ & + I_{[\tau,\infty)}(z_t)(C_2 + \Phi_{21}Y_{t-1} + \cdots + \Phi_{2p}Y_{t-p}) + \epsilon_t\,, \end{aligned}$$

where $Y_t = (y_{1t}, \dots, y_{nt})'$ is a $(n \times 1)$ vector, C_1 is a $(n \times 1)$ vector, Φ_{ji}, $j = 1, 2$, $i = 1, \dots, p$ are $(n \times n)$ matrices, and $\epsilon_t = (\epsilon_{1t}, \dots, \epsilon_{nt})$ is a vector of error terms with mean zero and positive definite covariance matrix Σ.

The TAR model has a wide range of applications. Tiao and Tsay [91], Potter [73], Pesaran and Potter [71], Rothman [78], and Koop and Potter [54] demonstrate both statistically significant and economically important nonlinearities in the US business cycle. Pfann, Schotman, and Tschernig [72] find strong evidence of high volatility and low volatility regimes in the behavior of US short-term interest rates. Dwyer, Locke, and Yu [26], Martens, Kofman, and Vorst [66], and Forbes, Kalb, and Kofman [33] describe the relationship between spot and futures prices of the S&P 500 index and model financial arbitrage in these markets as a threshold process. Obstfeld and Taylor [68] study the law of one price and purchasing power parity convergences and find strong evidence of two regimes. They demonstrate fast, months rather than years, convergence when price differences are higher than transaction costs, and slow or no convergence otherwise.

To simplify the exposition, our discussion of inference for all models will be conditional on the initial observations in the sample. We assume that $y_{1-p}, \ldots, y_0$ are observable. Two alternative treatments are possible. One can treat the initial observations as unobserved random variables and include the marginal density of initial observations into the likelihood. Alternatively, in the Bayesian analysis, one can treat the initial observations as any other parameter and augment the parameter space, θ, with $y_{1-p}, \ldots, y_0$.

Prior

The first step in Bayesian inference is to formalize prior beliefs about the model's parameters by choosing functional forms and parameters of prior distributions.

The prior density for τ depends on our choice of z_t. First, we can limit the prior support by the minimum and the maximum of z_t. Second, if $z_t = t$ the threshold is a date, and so the prior density is naturally discrete. If, however, $z_t = y_{t-d}$, the threshold τ is continuous and so is the prior density.

For a model to be identified, we restrict the support of the prior density to the region where the number of observations per regime is greater than the number of regressors. We assign an equal weight to the entire support to get the 'non-informative' prior for τ that is proportional to a constant

$$\pi(\tau) \propto I_{[z_{(k_1)}, z_{(T-k_2)}]}(\tau) , \tag{3}$$

where k_1 and k_2 are the number of regressors in the first and second regimes, and the subscript (t) indicates the order in the sample, $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(T)}$. For example, $z_{(1)} = 1$ and $z_{(T)} = T$ if z_t is a time index since the ordering is natural. For an alternative prior distribution of τ see Ferreira [31].

We assume that the prior density for β and σ^2 is independent of the prior density for τ. Also, because, conditional on τ, the model (2) is linear, we use the natural conjugate prior for β and σ^2

$$\begin{aligned} \pi(\beta|\sigma^2) &= N(\beta|\beta_0, \sigma^2 M_0^{-1}), \\ \pi(\sigma^2) &= IG_2(\sigma^2|\nu_0, s_0), \end{aligned}$$

where $IG_2(.)$ denotes the density of the Inverted Gamma-2 distribution. The functional form of the Inverted Gamma-2 density is given by

$$IG_2(\sigma^2|\nu, s) = \Gamma\left(\frac{\nu}{2}\right)^{-1} \left(\frac{s}{2}\right)^{\frac{\nu}{2}} \left(\sigma^2\right)^{-\frac{1}{2}(\nu+2)} \exp\left(-\frac{s}{2\sigma^2}\right) .$$

The natural conjugate prior allows us to use analytical integration that considerably simplifies the inference.

Estimation

The next step of the Bayesian analysis is to combine sample information with our prior beliefs to form the posterior beliefs. Given prior distributions, we update prior distributions with the sample likelihood into posterior distributions using Bayes' theorem. The posterior distribution can be further summarized for each parameter with its marginal expectation and variance.

Using the assumption of Normal errors, the likelihood function of the model (2) is

$$f(\beta, \sigma^2, \tau|y) \propto \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2} \sum (y_t - x_t'(\tau)\beta)^2\right\} . \tag{4}$$

The posterior density is a product of the prior and the likelihood

$$p(\beta, \sigma^2, \tau|y) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\tau) f(\beta, \sigma^2, \tau|y) . \tag{5}$$

Conditional on the threshold parameter, model (2) is linear. Applying the results from the standard natural conjugate analysis in the linear regression model (for details see Zellner [100]), the posteriors density of β, conditional on threshold and the data, can be obtained by integrating the posterior with respect to σ^2

$$\begin{aligned} p(\beta|\tau, y) &= \int p(\beta, \sigma^2|\tau, y) \, d\sigma^2 \\ &= t(\beta|\beta(\tau), s(\tau), M(\tau), \nu) , \end{aligned} \tag{6}$$

where $t(.)$ denotes the density of the multivariate Student t-distribution with

$$\begin{aligned} M(\tau) &= M_0 + \sum x_t(\tau)' x_t(\tau), \\ \beta(\tau) &= M(\tau)^{-1} \left(\sum x_t(\tau) y_t + M_0 \beta_0 \right), \\ s(\tau) &= s_0 + \beta_0' M_0 \beta_0 + \sum y_t^2 - \beta'(\tau) M(\tau) \beta(\tau), \\ \nu &= \nu_0 + T\,. \end{aligned}$$

Further, by integrating Eq. (6) with respect to β, we obtain the marginal posterior density for τ, which is proportional to the inverse of the integrating constant of $t(\beta|\beta(\tau), s(\tau), M(\tau), \nu)$ times the threshold prior density

$$p(\tau|y) \propto s(\tau)^{-\nu/2} |M(\tau)|^{-1/2} \pi(\tau)\,. \qquad (7)$$

Though analytical integration of this function is not available, the fact that it is a univariate function defined on bounded support greatly simplifies the numerical integration.

By integrating numerically the posterior for β conditional on the threshold and the data, we find marginal posterior density for β

$$p(\beta|y) = \int p(\beta|\tau, y) p(\tau|y)\, \mathrm{d}\tau.$$

Finally, using analytical results for the expectation of the conditional density β, we can find the marginal moments of β by integrating only over τ

$$\begin{aligned} E(\beta|y) &= \int E(\beta|\tau, y) p(\tau|y)\, \mathrm{d}\tau, \\ \mathrm{Var}(\beta|y) &= \int \mathrm{Var}(\beta|\tau, y) p(\tau|y)\, \mathrm{d}\tau \\ &\quad + \int (E(\beta|\tau, y) - E(\beta|y))(E(\beta|\tau, y) \\ &\quad - E(\beta|y))' p(\tau|y)\, \mathrm{d}\tau. \end{aligned}$$

Similarly, applying the results from the standard natural conjugate analysis, we obtain the posterior density of σ^2 conditional on the threshold and the data. Then we integrate out τ numerically to get the marginal posterior density for σ^2

$$p(\sigma^2|y) = \int IG_2(\sigma^2|\nu, s(\tau)) p(\tau|y)\, \mathrm{d}\tau\,,$$

and the marginal moments $E(\sigma^2|y)$ and $\mathrm{Var}(\sigma^2|y)$.

Testing for Linearity and Model Selection

After estimating the TAR model, we might ask whether our data are best characterized by two regimes or a single regime? Model (2) becomes linear when both regimes have identical regression coefficients, so that the difference $\beta_1 - \beta_2 = \delta$ is zero. There are two methods to the null hypothesis test $H_0 : \delta = 0$. The first approach is the Bayesian equivalent of the F-test. Taking into account that β conditional on τ has a Student t-distribution and that the linear transformation of a Student random vector is also a Student, the quadratic transformation of δ

$$\xi(\delta|\tau, y) = (\delta - \delta(\tau))' M_{22.1}(\tau)(\delta - \delta(\tau)) \frac{T-k}{k_2 s(\tau)} \qquad (8)$$

has a Fisher distribution, where

$$M_{22.1}(\tau) = M_{22}(\tau) - M_{21}(\tau) M_{11}^{-1}(\tau) M_{12}\,,$$

and $\delta(\tau)$ is our estimate. $M(\tau)$ is partitioned by dividing β into β_1 and δ. The posterior 'p-value' of the Bayesian F-test gives the unconditional probability that $\xi(\delta|y)$ exceeds $\xi(\delta = 0|y)$. It can be computed numerically as

$$\begin{aligned} &\Pr(\xi(\delta) > \xi(\delta = 0)|y) \\ &\quad = \int F(\xi(\delta = 0|y), k_2, T-k) \cdot p(\tau|y)\, \mathrm{d}\tau\,, \qquad (9) \end{aligned}$$

where $F(\xi(\delta = 0|y), k_2, T-k)$ is the Fisher distribution function with k_2 and $T-k$ degrees of freedom. The null hypothesis is accepted if, for example, $\Pr(\xi(\delta) > \xi(\delta = 0)|y)$ is larger than 5%.

The second approach, the posterior odds, is more general, and can also be used to select the number of lags p, the delay parameter d, or the number of regimes r. Koop and Potter [55,56] advocate and illustrate this approach in the context of the TAR model. To choose between two competing models, m_1 with $\theta_1 = (\beta_1, \delta, \tau, \sigma^2)$ and m_2 with $\theta_2 = (\beta_1, 0, \tau, \sigma^2)$, we calculate the posterior odds ratio

$$po_{12} = \frac{f(y|m_1)\pi(m_1)}{f(y|m_2)\pi(m_2)}\,,$$

where $\pi(m_i)$ is the prior probability for the model i, and $f(y|m_i)$ is the marginal likelihood or marginal density of the sample. Since $f(y|m_i)$ is a normalizing constant of the posterior density, it can be calculated as

$$f(y|m_i) = \int f(y|\theta_i, m_i) \pi(\theta_i|m_i)\, \mathrm{d}\theta_i\,.$$

With a 'non-informative' prior that assigns equal weight to each model, the posterior odds reduces to the ratio of marginal likelihoods, or the Bayes factor. Again, applying the standard natural conjugate analysis of the linear regression model, the marginal likelihood for model i

is

$$f(y|m_i) = \int \frac{\Gamma\left(\frac{\nu(\tau_i|m_i)}{2}\right) s_0^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)\pi^{\frac{T}{2}}} s(\tau_i|m_i)^{-\frac{\nu(\tau_i|m_i)}{2}} \cdot \left(\frac{|M_0|}{|M(\tau_i|m_i)|}\right)^{\frac{1}{2}} \pi(\tau_i|m_i)\,\mathrm{d}\tau \;, \qquad (10)$$

which can be calculated numerically. The model with the highest marginal likelihood is preferred.

Smooth Transition Autoregressive Model

In some applications, imposing an abrupt transition between regimes might be undesirable. For example, if the initial estimate of output is slightly below the threshold, even a small upward revision will result in a substantial change of the forecast in the TAR model. Bacon and Watts [6], in a regression model context, and Chan and Tong [14], in the TAR model context, propose to make the transition between regimes smooth. Terasvirta [89] develops a modeling cycle for the STAR model that includes specification, estimation, and evaluation stages as in the Box and Jenkins [9] modeling cycle for the linear time series model.

In the STAR model, a smooth transition is imposed by replacing the indicator function in Eq. (1) by the cumulative distribution function

$$y_t = [1-F(\gamma(z_t-\tau))]x_t'\beta_1 + F(\gamma(z_t-\tau))x_t'\beta_2 + \epsilon_t \,. \quad \text{(1a)}$$

Terasvirta [89] uses the logistic function

$$F(\gamma(z_t-\tau)) = \frac{1}{1+\exp(-\gamma(z_t-\tau))} \;,$$

where $\gamma \in [0,\infty)$ determines the degree of smoothness. As γ increases, smoothness decreases. In the limit, as γ approaches infinity, $F(.)$ becomes an indicator function, with $F(\gamma(z_t-\tau)) \sim 1$ when $z_t \geq \tau$. We can rewrite Eq. (1a) as

$$y_t = x_t'(\gamma,\tau)\beta + \epsilon_t \;, \qquad \text{(2a)}$$

where $x_t'(\gamma,\tau) = (x_t', F(\gamma(z_t-\tau))x_t')$.

Note that the identification problem discussed for the TAR model does not occur in the STAR model. We cannot have fewer observations than regressors because we no longer classify observations into regimes. The new parameter γ, however, introduces a new identification problem. If $\gamma = 0$, the logistic function equals $\frac{1}{2}$ for any value of τ, so τ is not identified. Also $x_t'(\gamma,\tau)$ is perfectly collinear unless the two regimes have no common regressors. Perfect collinearity implies that δ is also not identified. As in the TAR model, we choose such prior densities that resolve the identification problem.

The baseline model can be extended in several directions. Generally, the transition function $F(.)$ is not limited to the logistic function. Any continuous, monotonically increasing function $F(.)$ with $F(-\infty) = 0$ and $F(\infty) = 1$ can be used. For example, the popular alternative to the logistic function is the exponential function

$$F(\gamma(z_t-\tau)) = 1 - \exp(-\gamma(z_t-\tau)^2) \,.$$

In the regression model context, Bacon and Watts [6] show that results are not sensitive to the choice of $F(.)$. As in the TAR model, we can increase the number of regimes either with a single transition variable

$$\begin{aligned} y_t = {} & x_t'\beta_1 + F(\gamma_1(z_t-\tau_1))x_t'(\beta_2-\beta_1) + \ldots \\ & + F(\gamma_r(z_t-\tau_r))x_t'(\beta_r-\beta_{r-1}) + \epsilon_t \;, \end{aligned}$$

or with a combination of transition variables

$$\begin{aligned} y_t = {} & [(1-F(\gamma_1(z_{1t}-\tau_1)))x_t'\beta_1 + F(\gamma_1(z_{1t}-\tau_1))x_t'\beta_2] \\ & \cdot [(1-F(\gamma_2(z_{2t}-\tau_2)))] \\ & + [(1-F(\gamma_1(z_{1t}-\tau_1)))x_t'\beta_3 \\ & + F(\gamma_1(z_{1t}-\tau_1))x_t'\beta_4] \cdot [F(\gamma_2(z_{2t}-\tau_2))] + \epsilon_t \,. \end{aligned}$$

See van Dijk and Franses [97] for a discussion of the multiple regime STAR model.

Also, we can treat the choice of number of lags p, delay d, or number of regimes r as an inference problem, adding p, d, and r to the vector of parameters in the model. In addition, we can allow the variance of the error term to change between regimes, or more generally, use an autoregressive conditional heteroscedasticity form as in Lundbergh and Terasvirta [64], or a stochastic volatility form as in Korenok and Radchenko [59].

Finally, similar to the TAR model, the univariate STAR model can be extended to model a vector of time series as in Granger and Swanson [42]. The n dimensional two-regime STAR model can be specified as

$$\begin{aligned} Y_t = {} & [1-F(\gamma(z_t-\tau))](C_1 + \Phi_{11}Y_{t-1} + \cdots + \Phi_{1p}Y_{t-p}) \\ & + F(\gamma(z_t-\tau))(C_2 + \Phi_{21}Y_{t-1} + \cdots + \Phi_{2p}Y_{t-p}) \\ & + \epsilon_t \;, \end{aligned}$$

where we use the same notation as in the multivariate TAR model.

Applications of the STAR model include models of the business cycles, real exchange rates, stock and futures prices, interest rates, and monetary policy. Terasvirta and Anderson [88] and van Dijk and Franses [97] demonstrate nonlinearities in the US business cycles. Skalin and

Terasvirta [82] find similar nonlinearities in Swedish business cycles. Michael, Nobay, and Peel [67], Sarantis [80], and Taylor, Peel, and Sarno [87] show that the real exchange rate nonlinearly depends on the size of the deviation from purchasing power parity; Lundbergh and Terasvirta [65] and Korenok and Radchenko [59] use the STAR model to fit the behavior of exchange rates inside a target zone. Taylor, van Dijk, Franses, and Lucas [86] describe the nonlinear relationship between spot and futures prices of the FTSE100 index. Anderson [1] uses the STAR model to study yield movements in the US Treasury Bill Market. Finally, Rothman, van Dijk, and Franses [79] find evidence of a nonlinear relationship between money and output; Weise [99] demonstrates that monetary policy has a stronger effect on output during recessions.

Prior

As in the TAR model, the natural conjugate priors for β and σ^2 facilitate analytical integration. Bauwens, Lubrano, and Richard [7] impose the identification at $\gamma = 0$ by modifying the prior density of β

$$\pi(\beta|\sigma^2, \gamma) = N(\beta|0, \sigma^2 M_0^{-1}(\gamma)) \,,$$

where, assuming prior independence between β_1 and δ, M_0 is defined as

$$M_0(\gamma) = \begin{pmatrix} M_{0,11} & 0 \\ 0 & M_{0,22}/\exp(\gamma) \end{pmatrix} .$$

As γ gets closer to zero, the prior variance falls, increasing precision around $\delta = 0$. The choice of $\delta = 0$ is consistent with the linear hypothesis, which can be formulated as either $\delta = 0$ or $\gamma = 0$. When γ is positive, prior precision about $\delta = 0$ decreases as variance rises, so more weight is given to the information in the sample. We keep the natural conjugate prior of σ^2 without modifications.

We do not modify the prior for the threshold parameter τ. When γ is large, the smooth transition function is close to the step transition function. Thus, we prefer to limit the prior to the region where the number of observations per regime is greater than the number of regressors to avoid the TAR identification problem.

The prior for the smoothness parameter, γ, cannot be 'non-informative' or flat. As $\gamma \to \infty$ the smooth transition function becomes a step transition with a strictly positive likelihood. This means that the marginal likelihood function of γ is not integrable. To avoid the integration problem, Bauwens, Lubrano, and Richard [7] use the truncated Cauchy density

$$\pi(\gamma) \propto (1 + \gamma^2)^{-1} I_{[0,\infty)}(\gamma) \,.$$

Estimation

Inference in the STAR model follows the TAR methodology, taking into account the additional parameter γ, and the new definitions of $M_0(\gamma)$ and $x_t(\tau, \gamma)$.

In particular, the likelihood function of model (2a) is

$$f(\beta, \sigma^2, \tau, \gamma|y) \propto \sigma^{-T} \exp\left\{ -\frac{1}{2\sigma^2} \sum (y_t - x_t'(\tau, \gamma)\beta)^2 \right\} \,, \quad \text{(4a)}$$

the posterior density is

$$p(\beta, \sigma^2, \tau, \gamma|y) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\tau)\pi(\gamma) \, f(\beta, \sigma^2, \tau, \gamma|y) \,, \quad \text{(5a)}$$

and the joint posterior density of τ and γ is proportional to the inverse of the integrating constant of the Student t-density $t(\beta|\beta(\tau, \gamma), s(\tau, \gamma), M(\tau, \gamma), \nu)$ times the prior densities for c and γ

$$p(\tau, \gamma|y) \propto |s(\tau, \gamma)|^{-(T-k)/2} |M(\tau, \gamma)|^{-1/2} \, \pi(\tau)\pi(\gamma) \,, \quad \text{(7a)}$$

where

$$\begin{aligned} M(\tau, \gamma) &= M_0(\gamma) + \sum x_t(\tau, \gamma)' x_t(\tau, \gamma) \,, \\ \beta(\tau, \gamma) &= M(\tau, \gamma)^{-1} \left(\sum x_t(\tau, \gamma) y_t + M_0(\gamma)\beta_0 \right) \,, \\ s(\tau, \gamma) &= s_0 + \beta_0' M_0(\gamma)\beta_0 \\ &\quad + \sum y_t^2 - \beta'(\tau, \gamma) M(\tau, \gamma) \beta(\tau, \gamma) \,, \\ \nu &= \nu_0 + T \,. \end{aligned}$$

This function is bivariate and can be integrated numerically with respect to τ and γ. Then, as in the TAR model, we use numerical integration to obtain marginal densities and moments for β and σ^2.

Compared to the TAR model, β_1 and β_2 cannot be interpreted as regression coefficients in regime 1 and regime 2. Smooth transition implies that the effect of change in x_t on y_t is a weighted average of two regimes with weights changing from one observation to the other.

Testing for Linearity and Model Selection

The STAR model becomes linear when either $\delta = 0$ or $\gamma = 0$. The test for $H_0 : \delta = 0$ is equivalent to the test in the TAR model. The quadratic transformation of δ

$$\begin{aligned} &\xi(\delta|\tau, \gamma, y) \\ &\quad = (\delta - \delta(\tau, \gamma))' M_{22.1}(\tau, \gamma)(\delta - \delta(\tau, \gamma)) \frac{T-k}{k_2 s(\tau, \gamma)} \,, \end{aligned} \quad \text{(8a)}$$

where

$$M_{22.1}(\tau,\gamma) = M_{22}(\tau,\gamma) - M_{21}(\tau,\gamma)M_{11}^{-1}(\tau,\gamma)M_{12}(\tau,\gamma),$$

has a Fisher distribution. We can find the posterior 'p-value' of the Bayesian F-test numerically as

$$\begin{aligned}&\Pr(\xi(\delta) > \xi(\delta=0)|y)\\ &\quad = \iint F(\xi(\delta=0|y), k_2, T-k)p(\tau,\gamma|y)\,\mathrm{d}\tau\,\mathrm{d}\gamma\,.\end{aligned} \tag{9a}$$

The null hypothesis is accepted, for example, if $\Pr(\xi(\delta) > \xi(\delta=0)|y)$ is larger than 5%.

The test for $H_0 : \gamma = 0$ can be conducted using the 95% highest posterior density interval (abbreviated, HPDI), defined as the smallest interval with 95% probability of γ to be in the interval

$$\max_h \mathrm{PDI}(h) = \left\{\gamma \middle| \int p(\tau,\gamma)\pi(\tau)\,\mathrm{d}\tau \geq h\right\},$$
$$\text{s.t.} \Pr(\mathrm{PDI}(h)) \geq 0.95.$$

The null hypothesis is accepted, for example, if $\gamma = 0$ is inside the 95% HPDI.

As in the TAR model, linearity tests and model selection can be conducted using posterior odds. In the STAR model, the marginal likelihood for model i is given by

$$\begin{aligned}f(y|m_i) = \iint &\frac{\Gamma\left(\frac{\nu(\tau_i,\gamma_i|m_i)}{2}\right) s_0^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)\pi^{\frac{T}{2}}} s(\tau_i,\gamma_i|m_i)^{-\frac{\nu(\tau_i,\gamma_i|m_i)}{2}}\\ &\cdot\left(\frac{|M_0|}{|M(\tau_i,\gamma_i|m_i)|}\right)^{\frac{1}{2}} \pi(\tau_i|m_i)\pi(\gamma_i|m_i)\,\mathrm{d}\tau_i\,\mathrm{d}\gamma_i,\end{aligned} \tag{10a}$$

which can be calculated numerically. The model with the highest marginal likelihood is preferred.

Markov-Switching Model

Unlike the threshold models, where the regime transition depends on a time index or on lagged values of y_t, the Markov-switching autoregressive model relies on a random variable, s_t. A Markov-switching regression was introduced in econometrics by Goldfeld and Quandt [40] and was extended to the Markov-switching autoregressive model by Hamilton [45].

As in the threshold models, we limit our baseline MAR model to two regimes that differ only in mean. The variance of the error term is constant. The number of lags p is determined by the model choice. The two-regime MAR model becomes

$$\begin{aligned}&(y_t - \mu_{s_t}) = \sum_{i=1}^{p} \phi_i(y_{t-i} - \mu_{s_{t-i}}) + \epsilon_t,\\ &\mu_{s_t} = \mu_0 \qquad \text{if} \quad s_t = 0 \quad \text{(first regime)},\\ &\mu_{s_t} = \mu_0 + \mu_1 \quad \text{if} \quad s_t = 1 \quad \text{(second regime)},\end{aligned} \tag{11}$$

where $\mu_{s_t} = \mu_0 + s_t\mu_1$. An unobserved discreet random variable s_t takes only integer values of 0 or 1. The transition probability $\Pr(s_t = j|s_{t-1} = i) = p_{ij}$ that state i will be followed by state j depends only on s_{t-1}, the first order Markov-switching process, with transition probability matrix

$$P = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix}.$$

Since we have only two possible regimes and $p_{i1} + p_{i2} = 1$, we estimate only two free parameters, the probabilities of remaining in the same regime p_{11} and p_{22}. We also assume that, conditional on previous history of states $s = (s_1, \ldots, s_T)'$, the transition probabilities are independent of other parameters and the data.

In general, we do not have a clear association between regimes and the state indicator. This introduces an identification problem when we change regime identifiers, 0 and 1, and accordingly change $\mu_0^* = \mu_0 + \mu_1$ and $\mu_1^* = -\mu_1$. For example, if $s_t = 0$ during recessions, then the long run average during recessions is μ_0 and the long-run average during expansions is $\mu_0 + \mu_1$. On the other hand, if $s_t = 0$ during expansions, then the long-run average during expansions is $\mu_0^* = \mu_0 + \mu_1$ and the long-run average during recessions is $\mu_0^* - \mu_1$ or $\mu_1^* = -\mu_1$.

The second identification problem occurs in the MAR model when $\mu_1 = 0$; the model becomes linear. In this case, the conditional mean $E(y_t|s_t = 0) = E(y_t|s_t = 1) = \mu_0$ is independent of the state realizations, s, and transition probability matrix, P. Neither s nor P are identified.

The baseline model can be extended in several directions. The Markov-switching component can be modified by increasing the number of regimes as in Calvet and Fisher [11] and Sims and Zha [81] or by increasing the order of the Markov-switching process so that s_t depends on $s_{t-1}, \ldots, s_{t-r}$. Both changes can be incorporated by increasing the number of states in the baseline model, as in Hamilton [46].

Diebold, Lee, and Weinbach [22], Filardo [32], and Peria [69] relax the assumption of time invariant Markov-switching by making the transition probabilities depend on lagged values of y_t. In most applications, however, relatively few transitions between regimes makes it difficult to estimate the transition probabilities and restricts model choice to two or three regimes with time-invariant probabilities.

The error term can be modified by introducing regime-switching for the variance of the error term as in Hamilton and Susmel [47], and Cai [10]; by relaxing the assumption of Gaussian density for the error term as in Dueker [25]; or by specifying a general Markov-switching moving average structure for the error term as in Billio, Monfort, and Robert [8].

Finally, the univariate Markov-switching model can be extended to a multivariate model. Diebold and Rudebusch [23] propose a model where a number of time series are driven by a common unobserved Markov-switching variable, the dynamic factor model. The dynamic factor model captures the fact that many economic series show similar changes in dynamic behavior during recessions. Krolzig [60] provides a detailed exposition of how the baseline model can be extended to the Markov-switching vector autoregressive model.

The applications of the MAR model include models of business cycles, interest rates, financial crises, portfolio diversification, options pricing, and changes in government policy. Hamilton [45], Filardo [32], Diebold and Rudebusch [23], Kim and Nelson [51], Kim and Piger [53], and Hamilton [48] find statistically significant evidence that expansionary and contractionary phases of the US business cycle are distinct. Hamilton [44], Cai [10], Garcia and Perron [35], Gray [43], Dueker [25], Smith [83], Hamilton [48], and Dai, Singleton, and Yang [18] describe dramatic changes in interest rate volatility associated with the OPEC oil shocks, the changes in the Federal Reserve operating procedures in 1979–1982, and the stock market crash of October 1987. Ang and Bekaert [3] show a similar increase in volatility in Germany during the reunification period. Jeanne and Masson [49] use the MAR model to describe the crisis of the European Monetary System in 1992–1993; Cerra and Saxena [13] find permanent losses in output after the Asian crisis. Ang and Bekaert [2] report that the correlation between international equity returns is higher during bear markets relative to bull markets. Radchenko [76] shows that gasoline prices respond faster to a permanent oil price change compared to a transitory change. Finally, Sims and Zha [81] document abrupt changes of shocks to US monetary policy, and Davig and Leeper [20] document the regime changes in fiscal policy.

Prior

As in the threshold models, the natural conjugate priors facilitate considerably the integration of the posterior density. Conditional on s_t, μ_0, and μ_1, the MAR model is linear

$$y_t(s_t) = x_t'(s_t)\tilde{\phi} + \epsilon_t \,, \tag{12}$$

where

$$\begin{aligned} y_t(s_t) &= y_t - \mu_{s_t} \,, \\ x_t'(s_t) &= (y_{t-1} - \mu_{s_{t-1}}, \ldots, y_{t-p} - \mu_{s_{t-p}}) \,, \end{aligned}$$

and $\tilde{\phi} = (\phi_1, \ldots, \phi_p)'$. For the regression coefficient $\tilde{\phi}$ and the variance of the error term σ^2, the natural conjugate prior is given by

$$\begin{aligned} \pi(\tilde{\phi}|\sigma^2) &= N(\tilde{\phi}|\tilde{\phi}_0, \sigma^2 M_{0,\phi}^{-1}) I_A(\tilde{\phi}), \\ \pi(\sigma^2) &= IG_2(\sigma^2|\nu_0, s_0) \,, \end{aligned}$$

where A is a region where the roots of polynomial $1 - \phi_1 L - \cdots - \phi_p L^p = 0$ lie outside the complex unit circle. This restriction imposes stationarity on $y_t(s_t)$.

Conditional on s_t and $\tilde{\phi}$, the MAR model is also linear

$$y_t(\tilde{\phi}) = x_t'(\tilde{\phi})\tilde{\mu} + \epsilon_t \,, \tag{13}$$

where

$$\begin{aligned} y_t(\tilde{\phi}) &= y_t - \sum_{i=1}^{p} \phi_i y_{t-p} \,, \\ x_t'(\tilde{\phi}) &= \left(1, s_t - \sum_{i=1}^{p} \phi_i s_{t-p}\right) \,, \end{aligned}$$

and $\tilde{\mu} = (\mu_0, \mu_1)'$. The natural conjugate prior for $\tilde{\mu}$ is

$$\pi(\tilde{\mu}) = N(\tilde{\mu}|\tilde{\mu}_0, M_{0,\mu}^{-1}) I_{(0,\infty)}(\mu_1) \,,$$

where the indicator function imposes an identification constraint. In particular, we constrain the mean of the second regime to be greater than the mean of the first regime and in this way fix the order of regimes. We also impose $\mu_1 \neq 0$.

Kim and Nelson [51] show that the natural conjugate prior for the vector of transition probabilities $\tilde{p} = (p_{11}, p_{22})'$ is

$$\pi(\tilde{p}) = B(p_{11}|\alpha_1, \beta_1) B(p_{22}|\alpha_2, \beta_2) \,,$$

where $B(.)$ denotes the density of Beta distribution defined on the interval $[0, 1]$.

Estimation

In the Bayesian approach, we add realizations of the vector of states to the model parameters: $\theta = (\mu_0, \mu_1, \phi_1, \ldots,$

$\phi_p, \sigma, p_{11}, p_{22}, s_1, \ldots, s_T)'$. Analytical or numerical integration of the posterior density $p(\theta|y)$, where θ is $p + 5 + T \times 1$, may be difficult.

Albert and Chib [4] developed inference methodology that overcomes the curse of dimensionality using Gibbs-sampling, a Markov chain Monte Carlo simulation method of integration. The technique was further refined by Kim and Nelson [50]. Monte Carlo integration takes random draws from the posterior density and, by averaging them, produces estimates of moments. In particular, Gibbs-sampling allows us to generate many draws $\theta^{(g)}, g = 1, \ldots, G$, from joint density of $p(\theta|y)$ using only conditional densities $p(\theta_i|\theta_{i \neq j}, y)$ either for all i or for blocks of parameters. The joint and marginal distribution of $\theta^{(g)}$ converge at an exponential rate to the joint and marginal distribution of θ under fairly weak conditions. Casella and George [12], Gelfand and Smith [36], and Geweke [38] provide the details.

To implement the Gibbs-sampling simulation, we have to describe the conditional posterior distributions for all parameters or parameter blocks. It is convenient to separate parameters into five blocks: the state vector s, the transition probabilities $\tilde{p}$, the regression coefficients $\tilde{\phi}$ in the conditional linear model (12), the regression coefficients $\tilde{\mu}$ in the conditional linear model (13), and the variance of the error term σ^2.

The state vector s is a first-order Markov process, which implies that given s_{t+1} all information, for example $s_{t+2}, \ldots, s_T$ and $y_{t+1}, \ldots, y_T$, is irrelevant in describing s_t. Then the posterior density of s conditional on other parameters becomes

$$\begin{aligned} & p(s|\tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y) \\ & \quad = p(s_T|\tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y) \prod_{t=1}^{T-1} p(s_t|s_{t+1}, \tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y^t) \,, \end{aligned} \tag{14}$$

where $y^t = (y_1, \ldots, y_t)'$. The functional form of the posterior density suggests that we can generate draw of the state vector recursively. First we generate the last element s_T. Then, conditional on s_T, we generate s_{T-1}. More generally, conditional on s_{t+1}, we generate s_t for $t = T - 1, T - 2, \ldots, 1$.

To generate the state vector, Kim and Nelson [50] use the output from Hamilton's [45] filter. To facilitate exposition, we suppress the conditioning on parameters and consider first a model without lags.

Hamilton's filter starts from the observation that, before observing the data, the probability of finding the state in regime j, $\Pr(s_0 = j|y^0)$, equals the unconditional probability, $\Pr(s_t = j)$, which is proportional to the eigenvector of P associated with unitary eigenvalue.

Using transition probabilities and the probability of observing regime j conditional on observations obtained through date t, $\Pr(s_t = j|y^t)$, we predict the next period regime

$$\Pr(s_{t+1} = j|y^t) = \Pr(s_t = 0|y^t) p_{0j} + \Pr(s_t = 1|y^t) p_{1j} \,. \tag{15}$$

Once y_{t+1} is observed, we update the prediction using Bayes rule

$$\begin{aligned} \Pr(s_{t+1} = j|y^{t+1}) &= \Pr(s_{t+1} = j|y_{t+1}, y^t) \\ &= \frac{f(y_{t+1}|s_{t+1} = j, y^t) \Pr(s_{t+1} = j|y^t)}{f(y_{t+1}|y^t)} \,, \end{aligned} \tag{16}$$

where the numerator is the joint probability of observing y_{t+1} and $s_{t+1} = j$, which is a product of the probability of observing y_{t+1} given that state s_{t+1} is in regime j (for example $f(y_{t+1}|s_{t+1} = 0, y^t) = N(\mu_0, \sigma^2)$) and our prediction from Eq. (15). The denominator is the unconditional density of observing y_{t+1}, which is a sum of the numerator over all possible regimes

$$f(y_{t+1}|y^t) = \sum_j f(y_{t+1}|s_{t+1} = j, y^t) \Pr(s_{t+1} = j|y^t) \,. \tag{17}$$

Starting from $\Pr(s_0 = j|y^0)$, the filter iterates through Eqs. (15)–(17) until we calculate $\Pr(s_t = j|y^t)$ for every t and j. As a by-product of the filter we obtain the likelihood function

$$f(\tilde{\phi}, \tilde{\mu}, \tilde{p}, \sigma^2, s|y) = \prod_t f(y_{t+1}|y^t) \,. \tag{18}$$

For the AR(1) model, the filter should be adjusted. Given $\Pr(s_t = j|y^t)$, we forecast the next period regime and the previous period regime jointly, taking one summand in Eq. (15) at a time

$$\Pr(s_{t+1} = j, s_t = i|y^t) = p_{ij} \Pr(s_t = i|y^t) \,, \tag{15a}$$

for $j = 0, 1$ and $i = 0, 1$. After y_{t+1} is observed, we update our prediction to

$$\begin{aligned} & \Pr(s_{t+1} = j, s_t = i|y^{t+1}) \\ & = \frac{f(y_{t+1}|s_{t+1} = j, s_t = i, y^t) \ \Pr(s_{t+1} = j, s_t = i|y^t)}{f(y_{t+1}|y^t)} \,, \end{aligned} \tag{16a}$$

where $f(y_{t+1}|s_{t+1} = j, s_t = i, y^t)$ is the density of observing $y_t + 1$ given that state $s_t + 1$ is in regime j and state s_t is in regime i (for example $f(y_{t+1}|s_{t+1} = 0, s_t = 0, y^t)$ $=N(\mu_0 + \phi_1(y_t - \mu_0), \sigma^2))$

$$f(y_{t+1}|y^t) = \sum_j \sum_i f(y_{t+1}|s_{t+1} = j, s_t = i, y^t) \cdot \Pr(s_{t+1} = j, s_t = i|y^t) . \quad (17a)$$

Summing (16a) over i,

$$\Pr(s_{t+1} = j|y^{t+1}) = \sum_i \Pr(s_{t+1} = j, s_t = i|y^{t+1}), \quad (19)$$

finishes the iteration. Iterating through Eqs. (15a)–(17a) and (19) we get $\Pr(s_t = j|y^t)$ for every t and j. The extension to a more general AR(p) model is similar.

The output of Hamilton's filter gives only the first term in the product (14), which is sufficient to generate s_T. To generate the other states s_t conditional on y^t and $s_t + 1$, $t = T - 1, T - 2, \ldots, 1$, we again use Bayes rule

$$\Pr(s_t = j|s_{t+1} = i, y^t) = \frac{p_{ji} \Pr(s_t = j|y^t)}{\sum_j p_{ji} \Pr(s_t = j|y^t)} , \quad (20)$$

where $\Pr(s_t = j|y^t)$ is the output from Hamilton's filter. Since s_t is a discrete random variable taking on values 0 and 1, we can generate it by drawing random numbers from uniform distribution between 0 and 1, and comparing them to $\Pr(s_t = 1|s_{t+1} = i, y^t)$.

Conditional on other parameters in the model, the likelihood function of transition probabilities reduces to a simple count n_{ij} of transitions from state i to state j

$$f(\tilde{p}|\tilde{\mu}, \tilde{\phi}, \sigma_2, s, y) = p_{11}^{n_{11}}(1 - p_{11})^{n_{12}} p_{22}^{n_{22}}(1 - p_{22})^{n_{21}} ,$$

which is the product of the independent beta distributions. The posterior distribution for the transition probabilities conditional on the other parameters is a product of independent beta distributions

$$p(\tilde{p}|\tilde{\phi}, \tilde{\mu}, \sigma^2, s, y) \\ = B(\alpha_1 + n_{11}, \beta_1 + n_{12}) \cdot B(\alpha_2 + n_{22}, \beta_2 + n_{21}) .$$

To derive posterior distributions for $\tilde{\phi}$, $\tilde{\mu}$, and σ^2 conditional on other parameters, we use standard results for a linear model with the natural conjugate priors. The natural conjugate priors are reviewed, for example, by Geweke [39], Koop [58], or Lancaster [61]. In particular, the conditional distribution of the regression coefficients is Normal

$$p(\tilde{\phi}|\tilde{p}, \tilde{\mu}, \sigma^2, s, y) \\ = N\left(\Sigma_\phi\left(\sigma^{-2} M_{0,\phi}\tilde{\phi}_0 + \sigma^{-2}\sum x_t(s)' y_t(s)\right), \Sigma_\phi\right) \\ \cdot I_A(\tilde{\phi}),$$

$$p(\tilde{\mu}|\tilde{p}, \tilde{\phi}, \sigma^2, s, y) \\ = N\left(\Sigma_\mu\left(M_{0,\mu}\tilde{\mu}_0 + \sigma^{-2}\sum x_t(\tilde{\phi})' y_t(\tilde{\phi})\right), \Sigma_\mu\right) \\ \cdot I_{(0,\infty)}(\mu_1),$$

where

$$\Sigma_\phi = \left(\sigma^{-2} M_{0,\phi} + \sigma^{-2}\sum x_t(s)' x_t(s)\right)^{-1} ,$$

$$\Sigma_\mu = \left(M_{0,\mu} + \sigma^{-2}\sum x_t(\tilde{\phi})' x_t(\tilde{\phi})\right)^{-1} .$$

The conditional distribution for the variance of error term is Inverted Gamma-2

$$p(\sigma^2|\tilde{p}, \tilde{\phi}, \tilde{\mu}, s, y) \\ = IG_2\left(s_0 + \sum(y_t(s_t) - x_t'(s_t)\tilde{\phi})^2, \nu_0 + T\right) .$$

Testing for Linearity and Model Selection

Given our prior, the linear model is not nested in the MAR model. To test against a linear model, we use the Bayes factor. We also use the Bayes factor to select the number of regimes and the number of lags.

The Bayes factor is a ratio of marginal likelihoods of the alternative models. To find the marginal likelihood, we need to integrate the product of the likelihood function and the prior density with respect to all parameters. Chib [16] shows that the marginal likelihood can be computed from the output of the Gibbs sampler requiring only that the integrating constants of the conditional posterior distributions be known. This requirement is satisfied for the natural conjugate priors.

From the Bayes's theorem it follows that the identity

$$f(y) = \frac{f(y|\theta)\pi(\theta)}{p(\theta|y)} ,$$

holds for any θ. The complete functional form of the numerator is given by the product of the likelihood (18) and the prior densities. Chib suggests evaluating the denominator, the posterior density, at the posterior mode θ^*. Then the posterior density at the posterior mode can be written as

$$p\left(\theta^*|y\right) = p\left(\tilde{\mu}^*|y\right) p\left(\tilde{\phi}^*|\tilde{\mu}^*, y\right) \\ \cdot p\left(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y\right) p\left(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*}\right) .$$

The first term

$$p\left(\tilde{\mu}^*|y\right) = \int p\left(\tilde{\mu}^*|\tilde{\phi}, \sigma^2, \tilde{p}, s, y\right) p\left(\tilde{\phi}, \sigma^2, \tilde{p}, s|y\right)\ \mathrm{d}\tilde{\phi}\ \mathrm{d}\sigma^2\ \mathrm{d}\tilde{p}\ \mathrm{d}s,$$

can be estimated by averaging over the full conditional density

$$\hat{p}\left(\tilde{\mu}^*|y\right) = G^{-1}\sum_{g=1}^{G} p\left(\tilde{\mu}^*|\tilde{\phi}^{(g)}, \sigma^{2(g)}, \tilde{p}^{(g)}, s^{(g)}, y\right) .$$

This estimate converges at an exponential rate to the true marginal distribution of $\tilde{\mu}$.

In the second term,

$$p\left(\tilde{\phi}|\tilde{\mu}^*, y\right) = \int p\left(\tilde{\phi}^*|\tilde{\mu}^*, \sigma^2, \tilde{p}, s, y\right) p\left(\sigma^2, \tilde{p}, s|\tilde{\mu}^*, y\right)\ \mathrm{d}\sigma^2\ \mathrm{d}\tilde{p}\ \mathrm{d}s ,$$

the complete conditional density of $\tilde{\phi}$ cannot be averaged directly because the Gibbs sampler does not provide draws conditional on $\tilde{\mu}^*$. We generate necessary draws by additional G iterations of the original Gibbs sampler, but instead of generating $\tilde{\mu}$ we set it equal to $\tilde{\mu}^*$. Then the estimate of the second term

$$\hat{p}\left(\tilde{\phi}^*|\tilde{\mu}^*, y\right) = G^{-1}\sum_{g=G+1}^{2G} p\left(\tilde{\phi}^*|\tilde{\mu}^*, \sigma^{2(g)}, \tilde{p}^{(g)}, s^{(g)}, y\right) ,$$

converges at an exponential rate to the true $p\left(\tilde{\phi}|\tilde{\mu}^*, y\right)$. Similarly, by generating additional draws from the Gibbs sampler we compute $\hat{p}\left(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y\right)$ and $\hat{p}(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*})$.

Substituting our estimate of posterior density into marginal likelihood results in

$$\begin{aligned}\ln f\left(y\right) = \ln f\left(y|\theta^*\right) + \ln \pi\left(\theta^*\right) - \ln \hat{p}\left(\tilde{\mu}^*|y\right)\\ - \ln \hat{p}\left(\tilde{\phi}^*|\tilde{\mu}^*, y\right) - \ln \hat{p}\left(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y\right)\\ - \ln \hat{p}\left(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*}\right) .\end{aligned}$$

The model with the highest marginal likelihood is preferred.

Future Directions

Given the large volume of evidence collected in the nonlinear time series, incorporating regime-switching policies and disturbances into general equilibrium models may lead to a better understanding of monetary and fiscal policies.

Over the years, the time series literature has collected substantial statistical evidence that output, unemployment, and interest rates in the US exhibit different behavior in recessions and expansions. Contrary to the real business cycle models in which short-run and long-run fluctuations have the same origin, the statistical evidence suggests that the forces that cause output to rise may be quite different from those that cause it to fall.

Also, many studies provide evidence that monetary and fiscal policies have changed substantially throughout US history. Taylor [85], Clarida, Gali, and Gertler [17], Romer and Romer [77], and Lubik and Schorfheide [62] show that, since the mid-1980s, the Fed reacted more forcefully to inflation. Favero and Monacelli [30] and Davig and Leeper [20] demonstrate that US fiscal policy has fluctuated frequently responding to wars, recessions, and more generally to the level of debt. Sims and Zha [81], after extensive comparison of 17 regime-switching structural VAR models, report that their best-fitting model requires nine regimes to incorporate the large shocks, for example, generated by the OPEC oil embargo or the Vietnam War. They conclude that, "It is time to abandon the idea that policy change is best modelled as a once-and-for-all, nonstochastic regime switch" (p. 56).

The research by Davig and Leeper [19,20,21] and Farmer, Waggoner, and Zha [27,28,29] show considerable promise in introducing nonlinear regime-switching components into dynamic stochastic general equilibrium models. For example, Davig and Leeper [20] estimate regime-switching rules for monetary policy and tax policy and incorporate them into the otherwise standard new-Keynesian model. Unlike expansionary fiscal policy in the fixed-regime model, fiscal expansion in the regime-switching model increases inflation and output.

Acknowledgments

The author is grateful to Bruce Mizrach, the Finance and Econometrics Section editor, as well as to David Harless, Carol S. Lehr, Ming Lo, Stan Radchenko, Philip Rothman, and Tara Sinclair for many helpful comments.

Bibliography

1. Anderson HM (1997) Transaction costs and nonlinear adjustment towards equilibrium in the US treasury bill market. Oxf Bull Econ Stat 59:465–484

2. Ang A, Bekaert G (2002) International asset allocation with regime shifts. Rev Financ Stud 15:1137–1187
3. Ang A, Bekaert G (2002) Regime switches in interest rates. J Bus Econ Stat 20:163–197
4. Albert JH, Chib S (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J Bus Econ Stat 11(1):1–15
5. Astatkie T, Watts DG, Watt WE (1997) Nested threshold autoregressive NeTAR models. Int J Forecast 13:105–116
6. Bacon DW, Watts DG (1971) Estimating the transition between intersecting straight lines. Biometrica 62:525–534
7. Bauwens L, Lubrano M, Richard JF (1999) Bayesian Inference in Dynamic Econometric Models. Oxford University Press, New York
8. Billio M, Monfort A, Robert CP (1999) Bayesian estimation of switching ARMA models. J Econ 93:229–255
9. Box GEP, Jenkins GM (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco
10. Cai J (1994) A Markov model of switching-regime ARCH. J Bus Econ Stat 12:309–316
11. Calvet L, Fisher A (2004) Regime switching and the estimation of multifractal processes. J Financ Econ 2:49–83
12. Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46:167–174
13. Cerra V, Saxena SC (2005) Did output recover from the Asian crisis? IMF Staff Papers 52:1–23
14. Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Time Ser Anal 7:178–190
15. Chen CWS, Lee JC (1995) Bayesian inference of threshold autoregressive models. J Time Ser Anal 16:483–492
16. Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321
17. Clarida R, Gali J, Gertler M (2000) Monetary policy rules and macroeconomic stability: evidence and some theory. Q J Econ 115:147–180
18. Dai Q Singleton KJ, Yang W (2007) Regime shifts in a dynamic term structure model of US treasury bonds. Rev Financ Stud 20(5):1669–1706
19. Davig T, Leeper E (2005) Generalizing the Taylor principle. National Bureau of Economic Research, Working Paper No 11874
20. Davig T, Leeper E (2006) Fluctuating macro policies and the fiscal theory. In: Acemoglu D, Rogoff, Woodford M (eds) NBER Macroeconomic Annual. MIT Press, Cambridge
21. Davig T, Leeper E (2007) Generalizing the Taylor principle. Am Econ Rev 97(3):607–635
22. Diebold FX, Lee JH, Weinbach GC (1994) Regime switching with time-varying transition probabilities. In: Hargreaves C (ed) Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford
23. Diebold FX, Rudebusch GD (1996) Measuring business cycles: a modern perspective. Rev Econ Stat 78:67–77
24. DeJong DN (1996) A Bayesian Search for Structural Breaks in US GNP. In: Fomby TB (ed) Advances in Econometrics: Bayesian Methods Applied to Time Series Data, vol 11, part B. JAI Press, Greenwich, Connecticut, pp 109–146
25. Dueker M (1997) Markov switching in GARCH processes and mean-reverting stock-market volatility. J Bus Econ Stat 15:26–34
26. Dwyer GP, Locke P, Yu W (1996) Index arbitrage and nonlinear dynamics between the S&P 500 futures and cash. Rev Financ Stud 9:301–332
27. Farmer RE, Waggoner DF, Zha T (2006) Indeterminacy in a forward looking regime switching model. NBER Working Paper No 12540
28. Farmer RE, Waggoner DF, Zha T (2006) Minimal state variable solutions to Markov-switching rational expectations models. (unpublished manuscript)
29. Farmer RE, Waggoner DF, Zha T (2007) Understanding the New-Keynesian model when monetary policy switches regimes. (unpublished manuscript)
30. Favero CA, Monacelli T (2005) Fiscal policy rules and regime (in)stability: evidence from the US Manuscript, IGIER
31. Ferreira PE (1975) A Bayesian analysis of a switching regression model: known number of regimes. J Am Stat Assoc 70:370–374
32. Filardo AJ (1994) Business cycle phases and their transitional dynamics. J Bus Econ Stat 12:299–308
33. Forbes CS, Kalb GRJ, Kofman P (1999) Bayesian arbitrage threshold analysis. J Bus Econ Stat 17:364–372
34. Franses PH, van Dijk D (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge
35. Garcia R, Perron P (1996) An analysis of real interest under regime shift. Rev Econ Stat 78:111–125
36. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85(410): 398–409
37. Geweke J, Terui N (1993) Bayesian threshold autoregressive models for nonlinear time series. J Time Ser Anal 14: 441–454
38. Geweke J (1999) Using simulation methods for Bayesian econometric models: inference, development and communication. Econ Rev 18:1–127
39. Geweke J (2005) Contemporary Bayesian Econometrics and Statistics. Wiley, Hoboken
40. Goldfeld SM, Quandt RE (1973) A Markov model for switching regressions. J Econ 1:3–16
41. Granger CWJ, Terasvirta T (1993) Modeling Nonlinear Economic Relationships. Oxford University Press, Oxford
42. Granger CWJ, Swanson NR (1996) Future developments in the study of cointegrated variables. Oxf Bull Econ Stat 58: 537–553
43. Gray SF (1996) Modeling the conditional distribution of interest rates as a regime-switching process. J Financ Econ 42:27–62
44. Hamilton JD (1988) Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates. J Econ Dyn Control 12:385–423
45. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57(2):357–384
46. Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton
47. Hamilton JD, Susmel R (1994) Autoregressive conditional heteroskedasticity and changes in regime. J Econ 64:207–333
48. Hamilton JD (2005) Whats real about the business cycle? Fed Reserve Bank St Louis Rev 87(4):435–452
49. Jeanne O, Masson P (2000) Currency crises, sunspots, and Markov- switching regimes. J Int Econ 50:327–350

50. Kim CJ, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime-switching. Rev Econ Stat 80(2):188–201
51. Kim CJ, Nelson CR (1999) Friedman's plucking model of business fluctuations: tests and estimates of permanent and transitory components. J Money Credit Bank 31:317–334
52. Kim CJ, Nelson CR (1999) State-Space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications. MIT Press, Cambridge
53. Kim CJ, Piger J (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1189–1211
54. Koop G, Potter SM (1999) Dynamic asymmetries in US unemployment. J Bus Econ Stat 17:198–312
55. Koop G, Potter SM (1999) Bayes factors and nonlinearity: evidence from economic time series. J Econ 88:251–281
56. Koop G, Potter SM (2000) Nonlinearity, structural breaks or outliers in economic time series? In: Barnett B, Johansen S (eds) Nonlinear Econometric Modeling in Time Series. Cambridge University Press, Cambridge
57. Koop G, Potter SM (2003) Bayesian analysis of endogenous delay threshold models. J Bus Econ Stat 21(1):93–103
58. Koop G (2003) Bayesian Econometrics. Wiley, Chichester
59. Korenok O, Radchenko R (2005) The smooth transition autoregressive target zone model with the Gaussian stochastic volatility and TGARCH error terms with applications. VCU Economics Department, No 0505
60. Krolzig HM (1997) Markov-Switching Vector Autoregressions: Modeling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin
61. Lancaster T (2004) An Introduction to Modern Bayesian Econometrics. Blackwell Publishing, Malden
62. Lubik TA, Schorfheide F (2004) Testing for indeterminacy: an application to US monetary policy. Am Econ Rev 94:190–217
63. Lubrano M (1999) Bayesian Analysis of Nonlinear Time Series Models with a threshold. In: Barnett WA, Hendry DF, Hylleberg S, Terasvirta T, Tjostheim D, Wurts A (eds) Nonlinear Econometric Modeling. Cambridge University Press, Cambridge
64. Lundbergh S, Terasvirta T (1998) Modelling economic high-frequency time series with STAR-GARCH models. Working Paper Series in Economics and Finance No. 291, Stockholm School of Economics
65. Lundbergh S, Terasvirta T (2006) A time series model for an exchange rate in a target zone with applications. J Econ 131:579–609
66. Martens M, Kofman P, Vorst ACF (1998) A threshold error correction for intraday futures and index returns. J Appl Econ 13:245–263
67. Michael P, Nobay AR, Peel DA (1997) Transaction costs and nonlinear adjustment in real exchange rates: an empirical investigation. J Political Econ 105:862–879
68. Obstfeld M, Taylor AM (1997) Nonlinear aspects of goods-market arbitrage and adjustment: Heckscher's commodity points revisited. J Japan Int Econ 11:441–479
69. Peria MSM (2002) A regime-switching approach to the study of speculative attacks: a focus on EMS crises. In: Hamilton JD, Raj B (eds) Advances in Markov-Switching Models. Physica-Verlag, Heidelberg
70. Perron P, Vogelsang TJ (1992) Nonstationarity and level shifts with an application to purchasing power parity. J Bus Econ Stat 10:301–320
71. Pesaran MH, Potter S (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
72. Pfann GA, Schotman PC, Tschernig R (1996) Nonlinear interest rate dynamics and the implications for the term structure. J Econ 74:149–176
73. Poter SM (1995) A nonlinear approach to US GNP. J Appl Econ 10:109–125
74. Potter SM (1999) Nonlinear time series modelling: an introduction. J Econ Surv 13:505–528
75. Quandt RE (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. J Am Stat Assoc 53:873–880
76. Radchenko S (2005) Lags in the response of gasoline prices to changes in crude oil prices: the role of short-term and long-term shocks. Energy Econ 27:573–602
77. Romer CD, Romer DH (2002) A rehabilitation of monetary policy in the 1950s. Am Econ Rev 92:121–127
78. Rothman P (1998) Forecasting asymmetric unemployment rates. Rev Econ Stat 80:164–168
79. Rothman P, van Dijk D, Franses PH (2001) A multivariate STAR analysis of the relationship between money and output. Macroecon Dyn 5:506–532
80. Sarantis N (1999) Modeling non-linearities in real effective exchange rates. J Int Money Finance 18:27–45
81. Sims C, Zha T (2006) Were there switches in US monetary policy? Am Econ Rev 96:54–81
82. Skalin J, Terasvirta T (1999) Another look at swedish business cycles. J Appl Econ 14:359–378
83. Smith DR (2002) Markov-switching and stochastic volatility diffusion models of short-term interest rates. J Bus Econ Stat 20:183–197
84. Steel M (2008) Bayesian Time Series Analysis. In: Durlauf S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd ed. Palgrave Macmillan, London
85. Taylor JB (1999) An historical analysis of monetary policy rules. In: Taylor JB (ed) Monetary Policy Rules, pp 319–341
86. Taylor N, van Dijk D, Franses PH, Lucas A (2000) SETS, arbitrage activity, and stock price dynamics. J Bank Finance 24: 1289–1306
87. Taylor MP, Peel DA, Sarno L (2001) Nonlinear mean-reversion in exchange rate rates: towards a solution to the purchasing power parity puzzles. Int Econ Rev 42:1015–1042
88. Terasvirta T, Anderson H (1992) Characterising nonlinearities in business cycles using smooth transition autoregressive models. J Appl Econom 7S:119–136
89. Terasvirta T (1994) Specification, estimation and evaluation of smooth transition autoregressive models. J Am Stat Assoc 89:208–219
90. Terasvirta T (1998) Modelling economic relationships with smooth transition regressions. In Handbook of Appl Economic Statistics. Marcel Dekker, New York, pp 507–552
91. Tiao CG, Tsay RS (1994) Some advances in non linear and adaptive modelling in time series. J Forecast 13:109–131
92. Tong H (1978) On a threshold model. In: Chan CH (ed) Pattern Recognition and Signal Processing. Sijthoff and Noordhoff, Amsterdam

93. Tong H (1983) Threshold Models in Non-Linear Time Series Analysis. Lecture Notes in Statistics, no 21. Springer, Heidelberg
94. Tong H, Lim KS (1980) Threshold autoregression, limit cycles and cyclical data. J Royal Stat Soc B 42:245–292
95. Tong H (1990) Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, Oxford
96. Tsay RS (1998) Testing and modeling multivariate threshold models. J Am Stat Assoc 93:1188–1202
97. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroecon Dyn 3:311–340
98. van Dijk D, Terasvirta T, Franses PH (2002) Smooth transition autoregressive models – a survey of recent developments. Econ Rev 21:1–47
99. Weise CL (1999) The Asymmetric effects of monetary policy. J Money Credit Bank 31:85–108
100. Zellner A (1971) An Introduction to Bayesian Inference in Econometrics. Wiley, New York

Business Policy and Strategy, System Dynamics Applications to

JAMES M. LYNEIS
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary

Business policy and strategy A firm's business strategy defines how and where it competes, and its approach to doing so. A business strategy typically specifies a firm's goals, the products and services offered and the markets served, and the basis for competing (price, service, quality, etc.). A strategy may also define the organization structure, systems and policies which implement the strategy. In addition, firm's will have systems and policies which focus on operations and functions, and are not truly "strategic" in nature. Nevertheless, these operational policies can be important in determining business performance.

Business dynamics Business dynamics is the study of how the structure of a business (or a part of the business), the policies it follows, and its interactions with the outside world (customers, competitors, suppliers) determine its performance over time. Business structure consists of feedback loops surrounding the stocks and flows of resources, customers, and competitive factors that cause change over time; business policies are important components of these feedback loops. Business dynamics is a means of determining the likely performance that will result from alternative business policies and strategies.

Definition of the Subject

System dynamics has long been applied to problems of business performance. These applications range from operational/functional performance to overall strategic performance. Beginning with its founding at MIT's Sloan School of Management in 1957, an important focus of research, teaching, and application has been on understanding why companies and markets exhibit cycles, or underperform competitors in terms of growth or profitability. The original publication in the field was Forrester's Industrial Dynamics [26], which not only laid the theoretical foundations for the field, but also provided an understanding of the causes of instability in supply chains. Since that initial work, research and application has been widespread. It has addressed the dynamics underlying instability in manufacturing and service organizations, the processes which encourage or inhibit growth, the dynamics of research organizations, and the causes of cost and schedule overruns on individual projects. It has been applied in many industries, from manufacturing to high-tech to financial services and utilities, both by academics and consultants. Business theory and applications are taught at many universities, including but not limited to MIT, London Business School and others in England, Bergen (Norway), Manheim and Stuttgart (Germany) (see [62,72] for more details). Business policy and strategy has and will continue to be one of the major application areas for system dynamics.

Introduction

Business strategy, sometimes called simply 'policy' or 'strategy', is primarily concerned with how and where firm's choose to compete. It includes such decisions as setting goals, selecting which products and services to offer in which markets, establishing the basis for competing (price, service, quality, etc.), determining the organization structure, systems and policies to accomplish the strategy, and designing policies for steering that strategy continually into the future. Academic and applied research on business strategy developed separately from system dynamics. That research, while widely disparate, has largely focused on static assessments and tools. For example, cross-sectional studies of many companies attempt to identify key differences that determine success or failure as a guide to management; "strategic frameworks" (for example, learning curves, growth share matrices, Porter's five forces [79,80]) assist managers in framing strategy and intuitively assessing performance over time; scenario planning helps managers visualize alternative futures; the resource-based view of the firm and core competencies [81] help managers identify how resources and capabilities determine the best way to compete. While these tools provide valuable insights and frameworks, they leave the connection between a firm's business strategy and the evolution of its performance over time to the intuition of managers –

while traditional business strategy addresses the starting point and the desired end point, and the mechanisms that might allow the firm to transition between the two, the ability of those mechanisms to achieve that transition, and the path for getting between the two, is left unanswered.

Academic and applied research on operational and functional performance has similarly developed separately from system dynamics. Although it is difficult to generalize, this research is again typically static in nature and/or focused on the detailed management of a part of the organization over a relatively short period of time (for example, optimization of production scheduling during a month, quarter or year; optimal inventory management during a quarter or year). While this detailed management is necessary for running a business, it often overlooks the longer run implications of the policies established to manage the business in the short run, and of the impacts of one part of the business on other parts.

In contrast, system dynamics addresses how structure (feedback loops, stocks and flows) and policies determine performance over time – how does the firm, or a part of the firm, get from its current state to some future state. Evolving from this structural theory, system dynamicists have studied why firms and industries exhibit instability and cycles, and why firms grow or decline. Two real examples of problematic behavior over time are shown in Fig. 1. The example on the left shows the pattern of orders for commercial jet aircraft – a system dynamicist would try to understand why the orders are cyclical, and what can be done to make them less so (or to take advantage of the cycles by forecasting that cyclicality); the example on the right shows market shares of major players in a recently deregulated telecom market – a system dynamicist, working for the incumbent telecom, would try to understand the causes of market share loss and what can be done to reverse that loss.

From its beginnings in the late 1950s, system dynamics has been used to progressively develop structural theories to explain instability in supply chains [26], cycles of growth [28], boom and bust in product sales [100], and cost and schedule overrun on projects [54], to mention just a few. While system dynamics has at times borrowed, or in some cases reinvented, concepts from business policy and strategy, this structural theory development has until recently evolved largely independently of traditional business research and practice. There is, however, a great deal of potential synergy between traditional business research, particularly strategy research, and system dynamics that is increasingly being exploited.

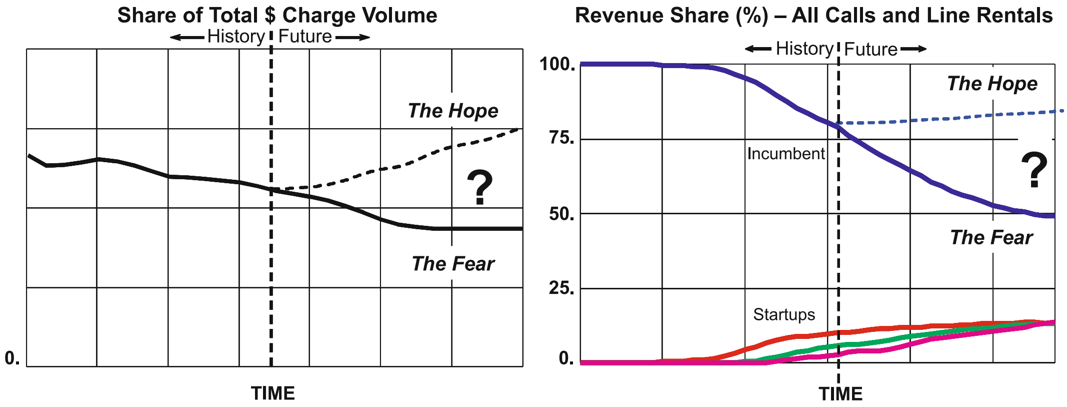
This paper surveys the application of system dynamics to business policy and strategy. The next section discusses the role of system dynamics models in policy and strategy formulation and implementation. Topics include how system dynamics fits into the typical policy and strategy formulation process, how system dynamics offers synergies with more traditional approaches, and how to conduct a modeling effort in order to enhance the implementation of any changes in policy or strategy. In Sect. Theory Development – Understanding the Drivers of Business Dynamics, system dynamics contribution to theory development is discussed – what are the structures underlying common business problems, and how can performance be improved? For example, what creates instability in supply chains, or "boom and bust" in product sales, and how can these behaviors be changed? Finally, in Sect. Applications and Case Examples applications to real world situations are presented – case studies that illustrate the value and impact of using system dynamics for policy and strategy development and implementation in specific firms and industries. As there has been a substantial amount of work done in this area, I must be selective, trying to touch on major themes and a representative sampling of work. Inevitably this will reflect my personal experiences, and my apologies to others that I have omitted either intentionally or unintentionally.

Using System Dynamics Models in Policy and Strategy Formulation and Implementation

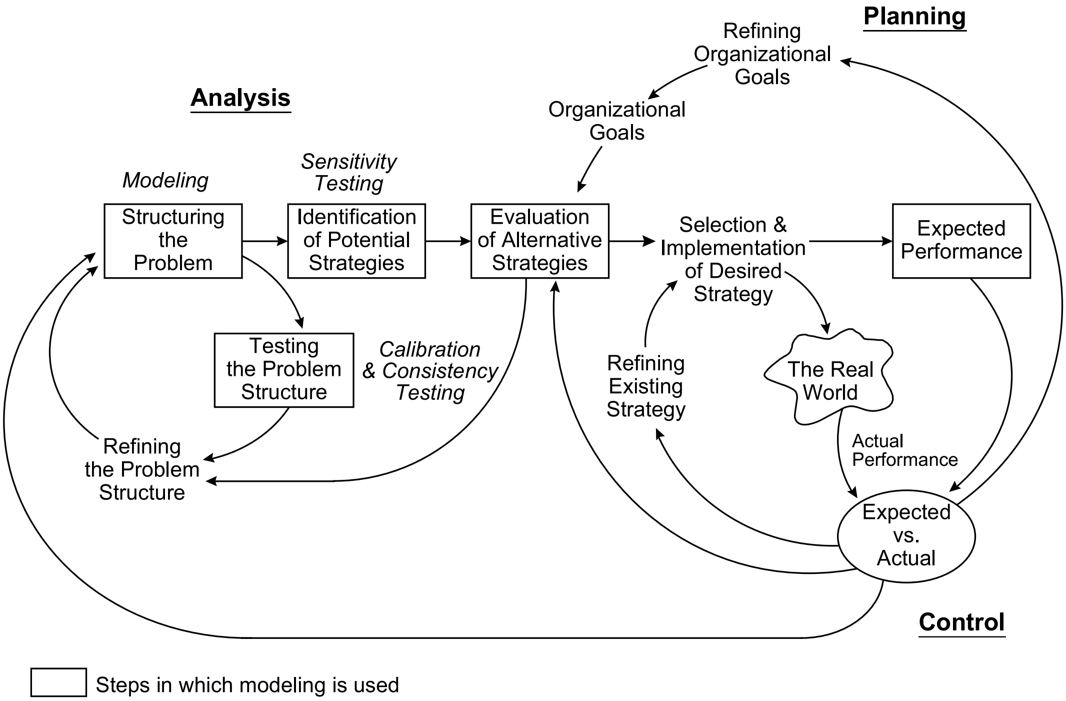
Role of System Dynamics Models in Policy and Strategy Formulation

There is general agreement among system dynamics modelers on the role that models play in the policy and strategy formulation process. This role has been depicted diagrammatically and described by Morecroft [65], Sterman [100], and Dyson et al. [23]. The role of the model in policy and strategy formulation is to act as a "virtual world" – a simpler and transparent version of the "real world". The model serves as a vehicle for testing our understanding of the causes of behavior in the real world, and as a laboratory for experimentation with alternative policies and/or strategies.

One version of that role is shown in Fig. 2. In many cases, the process starts with the definition of a problem – an aspect of behavior that is problematic or threatening. This might be a decline in market share or profitability, or the threat posed by a new competitive product or service, as illustrated in the example of Fig. 1. Sometimes the problem can be expressed in terms of achieving business goals or objectives in the future. As illustrated in Fig. 2, the overall policy/strategy management process can be divided into three components: analysis, planning, and con-



Business Policy and Strategy, System Dynamics Applications to, Figure 1
Examples of problematic behavior over time


Planning
Refining Organizational Goals
Analysis
Organizational Goals
Modeling
Sensitivity Testing
Structuring the Problem
Identification of Potential Strategies
Evaluation of Alternative Strategies
Selection & Implementation of Desired Strategy
Expected Performance
Testing the Problem Structure
Calibration & Consistency Testing
Refining Existing Strategy
The Real World
Actual Performance
Refining the Problem Structure
Expected vs. Actual
Control
Steps in which modeling is used

Business Policy and Strategy, System Dynamics Applications to, Figure 2
Policy/strategy management process and the role of system dynamics models [52]

trol. "Analysis" is usually triggered by a significant and/or persistent deviation between actual and expected performance. It involves the iterative structuring, testing and refinement of an organization's understanding of its operational or strategic problems and of the options open to it to deal with the performance gap. The model is the vehicle for this analysis – does our understanding of the system as reflected in model equations in fact produce behavior consistent with the observed problem, and if not, how can our understanding of structure be made more consistent with reality? The process of evaluating alternative policies/strategies often sheds new light on the problems faced by an organization or reveals the need for further analyzes. Note that the "modeling" cycle is iterative and compares simulated behavior to actual performance – the scientific method applied to strategy. Simulation of the model shows how business structure, policies, and external events together caused the past performance of the firm, and how future performance will evolve if structure, policies, and external events differ. The next phase, "planning", is also an iterative process; it involves the evaluation, selection, and implementation of policies/strategies – some authors refer to this as "rehearsing" strategy. Evaluation of alternative policies/strategies depends not only on projected accomplishment of organizational goals, but also on the realities of current performance. The existing operational policies or existing strategy (vision, mission, strategic objectives) and goals are subject to refinement, as required, based on the successes and problems encountered, and in response to changing conditions.

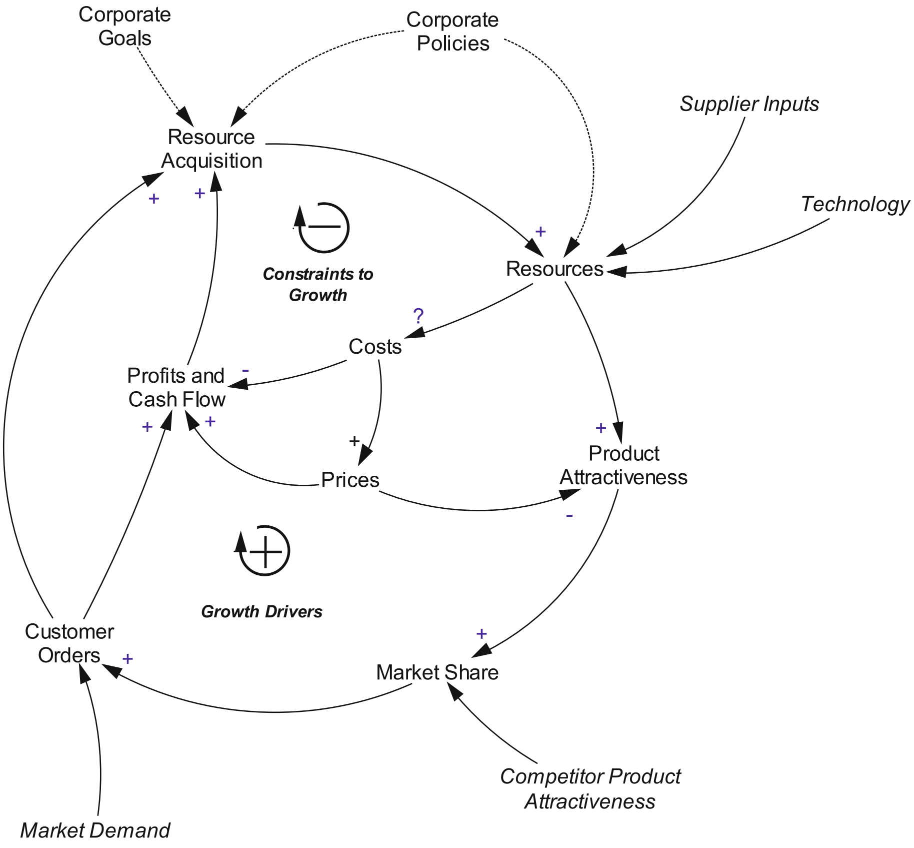
A third phase of the policy/strategy formulation process is here called "control". On-going policy/strategy *management* involves the continual, systematic monitoring of performance and the effective feeding back of successes, problems, threats, opportunities, experience, and lessons learned to the other components of the policy/strategy management process. The control phase is where organizations continue to learn. *The model provides an essential element to the control process – a forecast of expected performance against which actual performance can be monitored on a regular basis*. Deviations provide a signal for additional analysis: Has the policy/strategy been implemented effectively? Have conditions about the external environment changed? Are competitors acting differently than expected? Has the structure of the system changed? The model provides a means of assessing the likely causes of the deviation, and thereby provides an early warning of the need to act.

Synergy Between Traditional Strategy and System Dynamics

While Fig. 2 illustrates how system dynamics models fit into the policy/strategy formulation *process*, there is also a synergy between system dynamics models and traditional *strategy frameworks and concepts*. Figure 3 illustrates the factors that drive business performance from a system dynamics perspective. Starting with resources, a firm's resources determine its product attractiveness; a firm's market share is based on that attractiveness compared to the attractiveness of competitor products; market share drives customer orders, which in turn generates profits and cash flow to finance the acquisition of additional resources for further growth – thereby completing a growth-producing feedback around the outside of the figure (or as in the example of the telecom in Fig. 1, "growth"in the downward direction for the incumbent). However, the acquisition of additional resources can constrain future growth. To the extent increased resources increase costs, then profits and cash flow are reduced, and/or prices may need to increase. Both constrain growth (as might happen for the startups in the telecom example of Fig. 1).

There are a number of places in Fig. 3 where the system dynamics approach can be, and has been, connected to traditional strategy research and practice. For example, concepts such as learning curves, economics of scale, and economies of scope define possible connections between resources and costs – system dynamics models typically represent these connections. Figure 3 shows a number of factors external to the firm: market demand, competitor product attractiveness, technology, and supplier inputs. Strategy frameworks such as "five forces" and visioning approaches such as "scenario-based planning" [112], provide methods for thinking through these inputs – system dynamics models determine the consequences of alternative assumptions for the performance of the firm (note that system dynamics models also often internally represent structural dynamics of competitors, suppliers, and the market as appropriate to explain the behaviors and issues of interest, rather than specifying them as exogenous inputs). For example, Fig. 4 shows a "sector" diagram of the major components of a strategy model developed by a consulting company for a telecom company dealing with loss of market share as in Fig. 1 above (from [35], originally developed by Lyneis). The model not only represents factors internal to the dynamics of the telecom firm, but also factors related to the internal dynamics of competitors, regulatory responses to telecom and competitor performance, financial market responses to telecom performance, and market reactions to telecom and competitor competitive position. This sector diagram connects to a number of the traditional strategy frameworks. In addition to these general connections, a number of system dynamics papers have detailed specific connections to strategy concepts: learning curves and product portfolios [59]; duopoly competition [97]; diversification [31,68]; and industry structure and evolution [93].

Many of the connections between system dynamics and traditional strategy practice and research are discussed in Warren's *Competitive Strategy Dynamics* [105] and *Strategic Management Dynamics* [108]. More importantly, Warren's book details and expands on the connections between the system dynamics concepts of structure, particularly the concepts of stocks and flows, and the well-established resource-based view (RBV) of strategy and performance. (See [63], and [30], for explanations of the development of RBV; a managerial explanation of how this theoretical perspective can be applied can be found in Chap. 5 in [36]. Again, system dynamics provides a means



Business Policy and Strategy, System Dynamics Applications to, Figure 3
Drivers of business performance (adapted from [49])

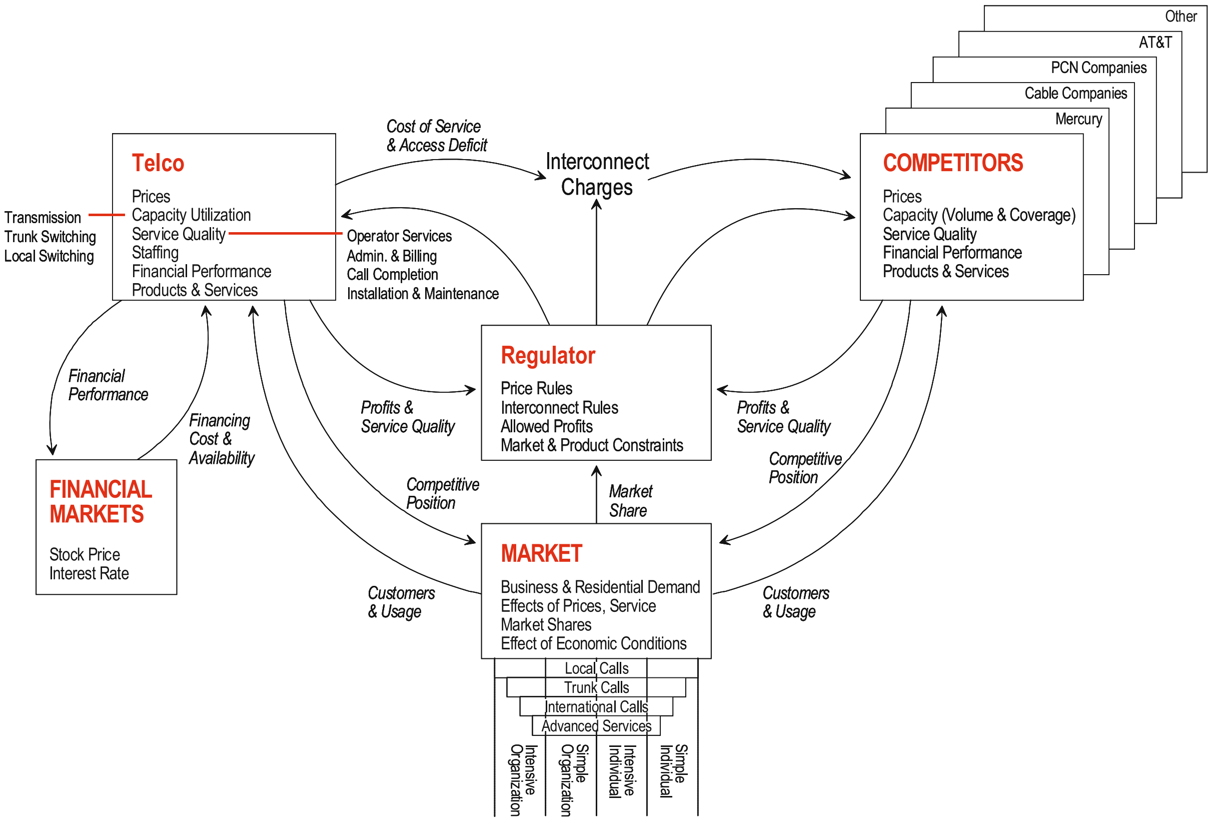
of simulating the consequences of alternative resource acquisition and allocation strategies on firm performance. As such, there would seem to be a strong synergy between system dynamics and this strategy approach.

In summary, while system dynamics and traditional strategy approaches developed largely independently, the potential synergies between the two are significant. Until recently, few researchers and practitioners have made the effort to cross disciplines and exploit this synergy. More effort and publication are needed to demonstrate areas of synergy and get system dynamics into the mainstream of business strategy research and ultimately practice.

Working with Management Teams to Achieve Implementation

While Fig. 2 depicts the overall role of the model in strategy management, the approach to developing and using the model itself involves an iterative, multi-phased process. That process has evolved over time as practitioners and researchers have learned from experience. Since its inception, system dynamics has been concerned with having an impact on business decisions. Jay Forrester, the founder of the field, stressed the importance of working on important problems – those that affect the success or failure of firms – and generating solutions that are relevant to those problems. Ed Roberts, one of the first researchers and practitioners in the field, was also involved in early research and experimentation on organizational change and how models can fit into that process [90,92]. The emphasis on having an impact has remained a central tenet of system dynamics, and over the years system dynamics practitioners have developed and refined methods of working with managers to not only solve problems, but also to enhance the likelihood of those solutions being implemented.

Starting with Roberts and his consulting firm Pugh-Roberts Associates (now a part of the PA Consulting Group), a number of researchers and practitioners have contributed to the evolving approach of working with managers to affect organizational change. These include,



Business Policy and Strategy, System Dynamics Applications to, Figure 4
"Sector" diagram from a typical strategy model

but are not limited to, Coyle and his group originally at Bradford University in the UK [14,15], Morecroft and his group at the London Business School [65,71], Richmond [88], Sterman [100] and Hines [37] at MIT, the "group model building" approach ([87,104]; ▶ Group Model Building and Peter Senge's organizational learning [95]. While there are some differences in emphasis and details, there is in general agreement on the high-level process of using system dynamics to affect corporate strategy development and change. In this section, I describe that process, discuss some specific areas where there is some divergence in practice, and end with some examples to the approach in practice.

In the early years, the approach of most system dynamicists to consulting was heavy on "product" and light on "process". Like management science in general, many in system dynamics took the view that as experts we would solve the client's problem for him, and present him with the solution. Practitioners gradually recognized that elegant solutions did not necessarily lead to implementation, and consulting styles changed to include increased client involvement [90,109]. At the same time, the "product" was evolving to meet the needs of clients (and to take advantage of the increased power of computers). Practitioners evolved from the smaller, policy-based models which characterized the original academic approach to more detailed models, along with the use of numerical time series data to calibrate these models and determine the expected numerical payoff to alternative strategies [52]. In addition, during the 1980s academic research began to focus more on "process:" the use of models in support of business strategy and on more effective ways to involve the client in the actual building of the model [65,66,87,88,104].

System dynamics practitioners now generally agree on a four-phased approach to accomplish these objectives:

1. Structuring the Problem
2. Developing an Initial Model and Generating Insight
3. Refining and Expanding the Model, and Developing a Strategy
4. On-going Strategy Management and Organizational Learning.

In practice, there is sometimes a fifth phase of work. Often modelers simplify a final project model in order to capture the core feedback loops that lie behind observed dynamics. Many of the generic structures discussed later in this paper arose in this way. In other cases, modelers will create management "games" and/or learning labs from the final project model.

As discussed below, the relative mix between "product" (detail complexity and calibration of model) and "process" (degree of involvement of client in model development and use) is perhaps the main difference in the style and approach of different practitioners of system dynamics. For those new to system dynamics, or for those seeking good examples, there is an excellent example of this process, including model development, by Kunc and Morecroft [44].

Phase 1 – Structuring the Problem The purpose of the first phase of analysis is to clearly define the problem of interest (either a past problem behavior or a desired future trajectory), the likely causes of that problem (or desired trajectory), and any constraints that may arise in implementing a solution. It identifies the performance objectives of the organization and possible solutions – all to be rigorously tested in later phases. Similar to more traditional policy and strategy approaches, during this phase the consultant team reviews company documents, the business press, and available company data, and interviews company managers, and possibly customers and competitors. During this phase, the hypothesized drivers of business performance are identified: what compels customers to buy this product, what compels them to buy from one supplier versus another, what drives the internal acquisition and allocation of resources, what major externalities affect the business (e. g., the economy, regulations, etc.), perhaps drawing on frameworks such as "five forces" and SWOT analyzes. More importantly, these drivers are linked in a cause-effect model to form a working hypothesis of the reasons for company behavior. This hypothesis formation builds heavily on the tools and techniques of what is now commonly called "systems thinking":

1. Behavior-over-time graphs (reference modes) – Graphs of problematic behavior over time, often with objectives for future performance highlighted (using actual data where readily available).
2. Causal-loop and mixed causal, stock-flow diagramming as a diagrammatic hypothesis of the causes of problematic behavior.
3. System archetypes, or common generic problem behaviors and structures observed over and over again in different businesses, as a means of identifying structure (see for example [95] and [43]); and
4. Mental simulation – does the hypothesis embodied in the conceptual model seem capable of explaining the observed problem(s)? Mental simulation is also used to identify the possible impact of alternative courses of action.

Note that the exercise to this point, as commonly practiced, is almost entirely qualitative. Warren [105,108] introduces quantitative dimensions even in this phase.

Phase 2 – Developing an Initial Model and Generating Insight The power of system dynamics comes from building and analyzing formal computer models. This is best done in two steps. In the first, a small, insight-based model is developed to understand the dynamics of the business so as to generate insights into the direction of actions needed to improve behavior. The small, insight-based model is also the next logical progression beyond "systems thinking" in the education of the client in the methods and techniques of system dynamics modeling. In the second quantitative modeling step (Phase 3 below), a more detailed version of the first model is developed, and is often calibrated to historical data. Its purpose is to quantify the actions needed, to assure that the model accurately reflects all relevant knowledge, and to sell others.

Small models (anywhere from 20–100 equations) make it much easier to understand the relationship between structure and behavior: how is it that a particular set of positive feedback loops, negative feedback loops, and stocks and delays interact to create the behavior shown in the simulation output? This can only be determined by experimentation and analysis, which is very difficult with large models. The focus of the first model is on insight generation, communication, and learning, rather than determining a specific shift in strategic direction or investment. These models can help managers improve their intuition (mental models) about the nature of their business, and thereby to better understand the rationale behind more detailed strategies that evolve in later phases of model development.

In summary, Phase 2 delivers:

- A small model which recreates the observed pattern of behavior or hypothesized future behavior (and is roughly right quantitatively);
- Analysis and understanding of the principal causes of that pattern of behavior;
- Ideas of high leverage areas that could improve behavior into the future; and

- Recommendations as to where additional detail will improve the strategy advice, or will make the results of the model more usable and/or easier to accept by others.

Phase 3 – Refining and Expanding the Model, and Developing a Strategy The final phase of model development entails the iterative expansion of the model to include more detail, and often calibration to historical data, as deemed appropriate for the situation. One progressively adds detail and structure, initially to make the process manageable, and then as necessary to correct discrepancies between simulated output and data, or to add policy handles and implementation constraints. Further, model development is likely to continue in the "on-going learning" phase as additional structure and/or detail is required to address new issues that arise. The purpose of this more elaborate modeling phase is to:

1. *Assure that the model contains all of the structure necessary to create the problem behavior*. Conceptual models, and even small, insight-based models, can miss dynamically important elements of structure, often because without data, the reference mode is incomplete or inaccurate (see [52] for examples of this).
2. *Accurately price out the cost-benefit of alternative choices*. Strategic moves often require big investments, and "worse-before-better" solutions. Knowing what is involved, and the magnitude of the risks and payoff, will make sticking with the strategy easier during implementation. Understanding the payoff and risks requires quantifying as accurately as possible the strengths of relationships.
3. *Facilitate strategy development and implementation*. Business operates at a detail level – information is often assembled at this level, and actions must be executed at that level. Therefore, the closer model information needs and results can be made to the normal business lines and planning systems of the company, the easier strategy development and implementation will be. And,
4. *Sell the results to those not on the client's project team*. Few, if any, managers can dictate change – most often, change requires consensus, cooperation, and action by others. The "selling" of results may be required for a number of reasons. If, as in the optimal client situation, consensus among key decision-makers is achieved because they are all a part of the project team, then the only "selling" may be to bring on board those whose cooperation is needed to implement the change. Under less optimal client circumstances where the project is executed by advisors to key decision-makers, or by a support function such as strategy or planning, then selling to decision maker(s) and to other functions will be required.

There are two important elements in this phase of work: adding detail to the model; and possibly calibrating it to historical data. Adding detail to the small, insight-based model usually involves some combination of: (1) disaggregation of products, staff, customers, etc.; (2) adding cause and effect relationships, and feedback loops, often where the more detailed disaggregation requires representing allocations, etc., but also to represent additional feedback effects that may seem secondary to understanding key dynamics, but may come into play under alternative scenarios, or may later help to "prove" the feedbacks were not important; (3) including important external inputs, typically representing the economy, regulatory changes, etc.; and (4) adding detailed financial sectors, which entail numerous equations, with important feedback from profitability and cash flow to ability to invest, employment levels, pricing, and so on. Calibration is the iterative process of adjusting model parameters, and revising structure, to achieve a better correspondence between simulated output and historical data. Whereas the Phase 2 model primarily relies on our store of knowledge and information about cause-effect structure, the Phase 3 model relies on our store of information about what actually happened over time.

In summary, Phase 3 delivers:

- An internally consistent data base of strategic information;
- A detailed, calibrated model of the business issue;
- A rigorous explanation and assessment of the causes of performance problems;
- Analyzes in support of strategic and/or tactical issues;
- Specific recommendations for actions; and
- Expectations regarding the performance of the business under the new strategy, and the most likely scenario.

Phase 4 – On-going Strategy Management System and Organizational Learning True strategy management ("control") involves the on-going, systematic monitoring of performance and the effective feeding back of successes, problems, threats, opportunities, experience, and lessons learned to the other components of the strategy management process. The control phase is where organizations continue to learn. *The model provides an essential element to the control process – a forecast of expected performance against which actual performance can be monitored on a regular basis*. Deviations provide a signal for additional analysis: Has the strategy been implemented effectively? Have conditions about the external environment changed? Are competitors acting differently than expected? Has the structure of the system changed? The model provides a means of assessing the likely causes of the deviation, and thereby provides an early warning of the need to act. This feedback is only possible with a detailed, calibrated model.

Differences in Emphasis and Style

While there is general agreement among system dynamics practitioners regarding the role of models in the strategy development process and of the basic steps in that process as described above, there are some differences in emphasis and style regarding: (1) the use of causal-loop diagrams (CLDs) vs. stock-flow (SF) diagrams; (2) whether you can stop after Phase 1 (i. e., after the "qualitative" phase of work); (3) is calibration necessary and/or cost effective; and (4) how much model detail is desirable.

CLDs vs. SF There are disagreements within the field about the value of causal loop diagramming (versus stock-flow diagrams). Causal-loop diagrams focus on the feedback loop structure that is believed to generate behavior; stock-flow diagrams also include key stocks and flows, and in the extreme correspond one-to-one with complete model equations. In my view, there is no "right" answer to this debate. The most important point is that in Phase 1 diagramming one is trying to develop a dynamic hypothesis that can explain the problem behavior, and that can form the basis of more detailed diagramming and modeling – whether that dynamic hypothesis is a CLD, a stock-flow diagram with links and loops labeled, or some combination depends in part on:

- Personal style and experience – some people, Jay Forrester perhaps being the best example, seem to always start with the key stocks and flows and work from there; Kim Warren [106] also argues for this approach as an effective means of connecting to the way managers view the problem and to the data;
- The structure of the system – some systems have "obvious" key chains of stocks and flows, and so starting there makes the most sense (for example, the aging chains in the urban dynamics model [29], the rework cycle on projects, and inventory control systems); other systems, without critical chains of stocks and flows, may be easier to address starting with CLDs;
- Whether or not you are doing the model for yourself or with a group – especially if the group is not conversant with the basics of system dynamics, starting with CLDs is easier, and it's also easier to brainstorm with CLDs (which is different than developing a dynamic hypothesis); but again, it's personal preference and nature of system as well. In practice, I have found that CLDs alone, or a mixed stock-flow/causal diagram, are extremely valuable for eliciting ideas in a group setting about the cause-effect structure of the business, and later for explaining the dynamics observed in simulation output. However, one cannot build a model literally from a causal diagram, and either explicit or implicit translation is required.

Qualitative vs. Quantitative Modeling Some practitioners of system dynamics believe that strategic insights can sometimes be obtained after the first phase of work, after the dynamic hypothesis and mental simulation (and note that much of traditional strategy practice relies on such qualitative insights). Coyle [17,18] argues for this; the popularity of "systems thinking" engendered by Senge's work [95] has spawned a number of practitioners that use only qualitative modeling [52]. Coyle's views generated a strong counter response from [41]. Wolstenholme [110] provides a history and discusses his view on the advantages and disadvantages of each approach. My own view is that while Phase 1 and the systems thinking that is a key part of it are a necessary start, it should not be the end point. Two problems limit its effectiveness in supporting business strategy. First, simple causal diagrams represented by system archetypes, while useful pedagogically, take a very narrow view of the situation (typically, one or two feedback loops). In reality, more factors are likely to affect performance, and it is therefore dangerous to draw policy conclusions from such a limited view of the system. A more complete representation of the problem considers more feedback effects and distinguishes stocks from flows, but introduces the second problem: research has shown that the human mind is incapable of drawing the correct dynamic insights from mental simulations on a system with more than two or three feedback loops [78,98]. In fact, without the rigor and check of a formal simulation model, a complex causal diagram might be used to argue any number of different conclusions. In addition to overcoming these limitations, as discussed below, formal modeling adds significant value to the development and implementation of effective business strategies. Warren (p. 347 in [107]) also stresses need to focus on quantitative behavior to achieve management consensus.

Need for Data/Validation The necessity of obtaining numerical data and calibrating model output to that data is also questioned by some practitioners. While I agree that curve fitting via exogenous variables is not a useful endeavor, proper calibration is an important part of the scientific method that involves systematically comparing simulation output to data, identifying causes of error, and correcting discrepancies by improving first the structure of the model and then its parametrization. In some cases, discrepancies are ignored because they are deemed to be caused by factors irrelevant to the problem of interest, or may be "fixed" by exogenous factors if these are deemed significant by the client and are consistent with the remaining model structure and calibration. As Homer [38,39] argues, the use of historical data and calibration is essential to scientific modeling.

In some cases, organizations lack the data on key factors felt to be essential to the dynamic performance of the business, and by implication essential to sound strategic management of the business. The modeling process can highlight these short-comings and, in the short-term, substitute educated assumptions for this data. In the longer-term, companies can be encouraged to acquire this important data (and substitute it for much of the unimportant information companies generally pore over).

Accurate calibration can greatly enhance confidence in a model. This can be especially important when trying to convince others of the appropriateness of actions a management team is going to take, or to demonstrate to others the need to take action themselves based on the results of the model. Calibration can also be important for other reasons: (1) numerical accuracy is often necessary to evaluate the relative cost and benefits of changes in strategy, or to assess short-term costs before improvements occur; (2) calibration often uncovers errors in the data or other models, especially incomplete or incorrect mental models that form the basis for the dynamic hypothesis (see [52] for examples); and (3) the "control" feedback in the fourth phase of the strategy management process is only possible with a detailed, calibrated model.

Level of Detail and Model Complexity Some practitioners argue that large, complex models should be avoided, for a number of reasons: they can be even more like black boxes; they can be difficult to understand (not only for the non-modelers, but even the modelers); and they are costly to develop. Morecroft argues that a detailed model "loses its agility and becomes less effective as a basis for argument". (p. 227 in [65]) In practice, the first two issues can be avoided and/or minimized by executing the model development in three phases as discussed above. This allows the client to grow slowly with the concepts, and it allows the modeling team to develop a solid understanding of the model. The third problem is generally not an issue if you are working on significant problems – in my view the cost of the consulting engagement is trivial relative to the expected payoff. While I believe that the client obtains value, regardless of when you stop, strategy consulting is one case where the "80/20 rule" does not apply – the client does not get 80% of the value for 20% of the cost (which would be essentially at the end of Phase 1). In part this is a function of what I view as the objective of the project – providing tools, strategic analyzes, and advice in support of an important strategic and/or investment decision. In this situation, the "value" is back-end loaded. Finally, effective strategy management is only possible with a detailed, calibrated model.

In addition, detail and calibration are often necessary to sell the model to others. In many situations, everyone who may have an input to a strategic decision or be necessary for successful implementation cannot be a part of the client team. As surprising as it may seem, the selling of results (as opposed to understanding) is easier to accomplish with a detailed, calibrated model than with a small model. First, the numerical accuracy gives the model face validity. Second, a detailed model more often allows the modeler to counter the "have you considered (insert pet theory)?" criticism. I have often found that when you start explaining the model to others, they respond by asking "Have you considered this feedback? Or this effect?" And if you have not, that ends the discussion. Even though you may think that feedback or that effect may not have any impact, if it is not included in the model you cannot say "Yes, we looked at that and it did not have any impact", and explain why. If it is not in the model the critic can argue that your results would be changed by the inclusion of their pet theory. One has a hard time countering that assertion without a convincing argument based on simulation results. Finally, a detailed, calibrated model helps tell a convincing story. The simulation output, which corresponds closely to the data, can be used to explain (again with output) why, for example, a loss of market share occurred. How price relative to the competitions' price was the key factor, and/or how the factors affecting share changed over time. The simulation output can and should be tied to specific events. We have found that an explanation like this is compelling, and is important in enhancing the credibility of the model and the modeler.

The benefits of large, complex models in a consulting setting are also noted by Winch [111]. He specifically finds that "For the executive team to have confidence in the impartiality of the model, each person must feel it captures the detailed pressures and processes of his or her own sphere of responsibility yet produces a holistic view of the organization". (pp. 295–6 in [111]), and that the model was essential to getting everyone to agree: "The process of building system dynamics models, in each case ostensibly as a forecasting and evaluation tool, enabled the managers eventually to develop a shared view, which formed the basis for formulating and agreeing upon a final strategy". (p. 298).

Process Examples

There are a number of published examples that support the four-phase process of applying system dynamics to business strategy:

- Lyneis [52] provides not only a more fully developed description of the detailed, calibrated-model Pugh–Roberts approach, but also illustrates its application to the credit card and airline manufacturing industries.
- Morecroft et al. [70] describe how a model was created and used to stimulate debate and discussion about growth management in a biotechnology startup firm. The paper highlights several novel features about the *process* used for capturing management team knowledge. A heavy emphasis was placed on mapping the operating structure of the factory and distribution channels. Qualitative modeling methods (structural diagrams, descriptive variable names, "friendly" algebra) were used to capture the management team's descriptions of the business. Simulation scenarios were crafted to stimulated debate about strategic issues such as capacity allocation, capacity expansion, customer recruitment, customer retention, and market growth, and to engage the management team in using the computer to design strategic scenarios. The article concludes with comments on the impact of the project.
- Winch [111] examines the role that building and using a system dynamics model plays in developing consensus within management teams facing key strategic decisions: A shared view emerges within the team as individual views of the company, its industry, and the socioeconomic climate are articulated and compared. Examples are given based on two actual consulting assignments in which differing views concerning the competitive environment and the general business outlook initially pointed to quite different strategies. The emergence of consensus was considered a major benefit in addition to the forecasts and quantitative evaluations the model provided. In its analysis and examples, this article emphasizes both the "hard" benefits of forecasts and an objective framework for quantitative evaluations and the "soft" benefits of building consensus within management teams.
- Coyle [15,16] also has an approach that he discusses, with emphasis on CLDs (he terms these "influence diagrams", and his group was instrumental in initial use of this technique).
- Snabe and Grossler [94] show how modeling can be supportive for strategy implementation in organizations and illustrate with a detailed case study from a high-tech company.
- A special issue of the Journal of the Operational Research Society on System Dynamics for Policy, Strategy, and Management, edited by Coyle and Morecroft [19], contains a number of papers which in part discuss consulting process issues [21,110,112] among others).
- The special issue Fall 2001 of *System Dynamics Review* on consulting practice contains papers by Thompson [102], Campbell [11], and Backus et al. [5] that focus on the consulting process.

Theory Development – Understanding the Drivers of Business Dynamics

Another important contribution of system dynamics to business policy and strategy formulation is the development of structural theories to explain commonly observed patterns of behavior. Theory development provides us with: an understanding of the basic drivers of business dynamics; insights, enhanced mental models, and policy guidelines for improved performance; and building blocks of tested model equations for real applications (equations for the model must be provided for models to add to our base of theory). System dynamicists have developed structural theories to explain the basic patterns of business dynamics: (1) cycles and instability; (2) productivity and eroding performance; (3) life cycles; and (4) growth. Each is discussed in turn below.

Cycles and Instability: Stock Management, Supply Chains, and Manufacturing Systems

The very first applications of system dynamics were to understanding the tendencies of production-distribution systems, or "supply chains", toward cycles and instability; these applications remain important to this day [25,26]. For example, the "Beer Game", now distributed by the System Dynamics Society, was developed and refined at MIT beginning in the early 1960s and remains one of the most popular introductions to both system dynamics principles, and to supply chain issues.

Supply chains are an important component of all industrialized societies. They exist in any industry where goods are produced and distributed to consumers, for example, food and beverage production and distribution, or manufactured goods such as automobiles and appliances. Supply chains exhibit a classic behavior pattern which has impacts not only on the individual company, but also for the economy as a whole: as one moves up the supply chain from the end user, any variation in orders from the end user are progressively *amplified* and delayed at each additional stage in the chain –factory variations are greater than customer variations; raw materials production variations are greater than factory variations (see Chap. 17 and 20 in [100] for real world examples of this behavior). This behavior is also sometimes referred to as the "bullwhip"effect.

Figure 5 illustrates the structure of one stage of a typical supply chain and the causes of amplification: a stock of inventory is depleted by shipments (here assumed equal to demand) and replenished by production completions (or more generally, shipments from a supplier); the stock of goods in production (or goods being assembled and shipped by a supplier) are increased by production and reduced, after the production (and/or shipping) delay, by production completions. This structure has a tendency to "amplify" any changes in demand – that is, "production" (or orders and reorders, depending on the system) increase or decrease more than any increase or decrease in demand, and tend to lag changes in demand. For example, in Fig. 5, when demand increases, even if production increases immediately inventory falls because production completions are delayed by the production (and/or shipping) delay. Therefore, production must increase higher than demand (amplification) in order to rebuild inventories. In addition to production and shipping delays, inventory might also fall because of delays caused by smoothing information about demand (such that production changes lag changes in demand). Production further increases above demand because of the need to increase inventories and production or supply lines to higher target levels. Intuitively, and verified by simulations, amplification is greater if: desired inventories are larger; production/transit delays are longer; and/or responses to inventory gaps are more aggressive (smaller adjustment time constant, as discussed below).

This basic structure in Fig. 5 also illustrates the "stock management" problem. In Fig. 5, the stock of finished goods inventory must be managed in order to serve customer demand in a timely fashion. Figure 6 details the structure typically used to control stocks, one of the most used and important structures in system dynamics:

$$\begin{aligned}&\text{Production} = \text{Expected Demand}\\&\quad + \text{Inventory Correction} + \text{Goods In Process Correction}\end{aligned}$$

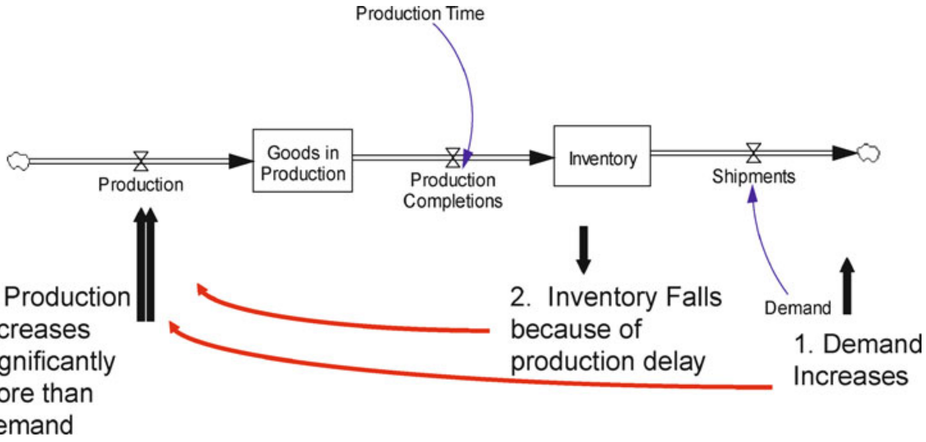
$$\begin{aligned}&\text{Inventory Correction}\\&\quad = (\text{Desired Inventory} - \text{Inventory})/\text{Time to Correct}\end{aligned}$$

$$\begin{aligned}&\text{Inventory Desired Inventory}\\&\quad = \text{Expected Demand} \times \text{Months Coverage Goal}\end{aligned}$$

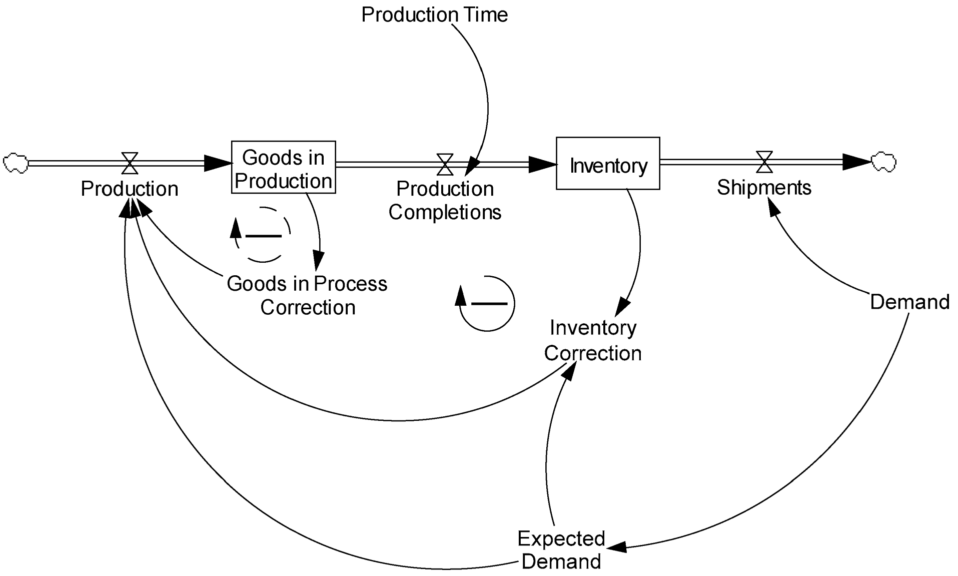
$$\begin{aligned}&\text{Goods In Process Correction} =\\&(\text{Desired Goods In Production} - \text{Goods In Production})\\&/\text{Time to Correct Inventory}\end{aligned}$$

$$\begin{aligned}&\text{Desired Goods In Production}\\&\quad = \text{Expected Demand} \times \text{Production Time}\end{aligned}$$

Stock management is complicated by delays in replenishing the stock, here a production delay. Depending on the pattern of demand, there is often a tradeoff between



Business Policy and Strategy, System Dynamics Applications to, Figure 5
Structure and causes of amplification in one stage of a supply chain



Business Policy and Strategy, System Dynamics Applications to, Figure 6
Stock management structure

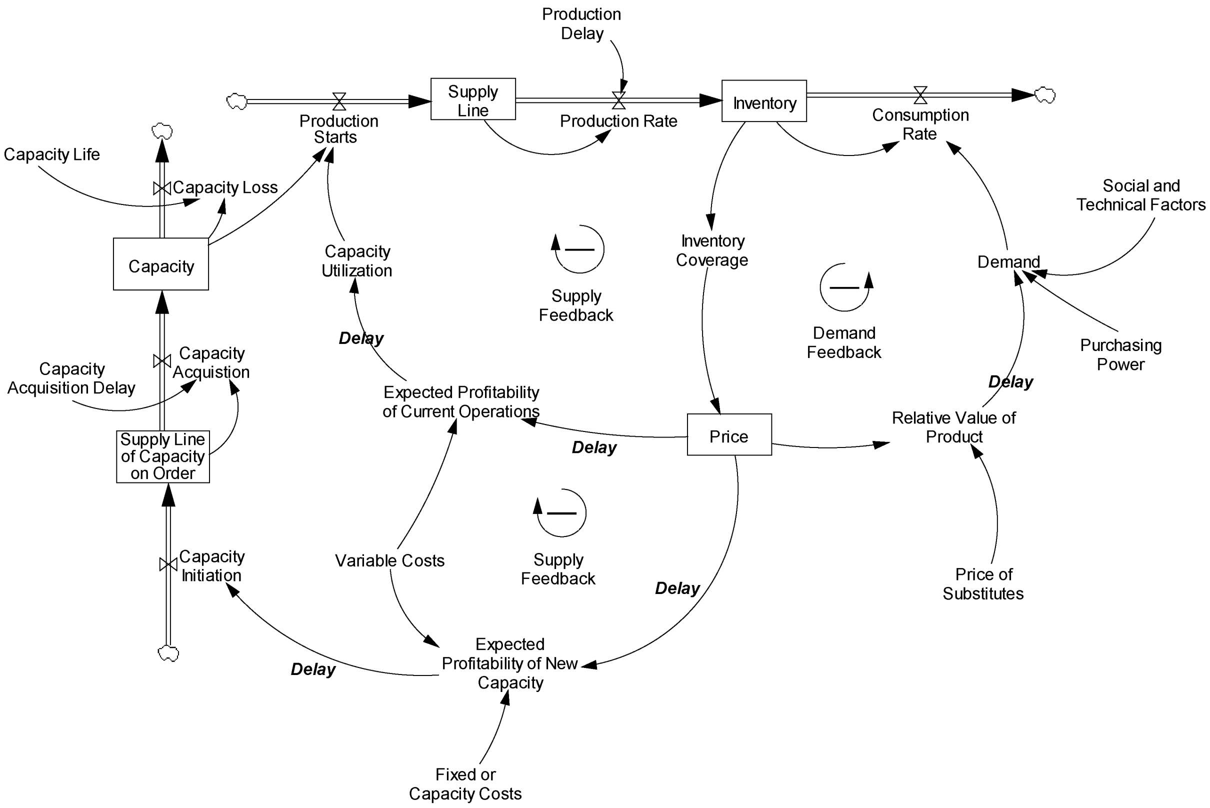
amplification and variations in inventory –less aggressive responses (longer time to correct inventory) generally reduce amplification but cause greater variations in inventory (and therefore may necessitate higher target levels to reduce the likelihood of stockouts); more aggressive responses (shorter time to correct inventory) increase amplification and demands on manufacturing and suppliers, and potentially costs, but can result in more stable inventory levels. However, under some demand patterns and production conditions, aggressive responses can increase both amplification and inventory instability. While as noted below structural changes can significantly improve the overall performance of stock management and supply chain systems, nevertheless this fundamental tradeoff between amplification and inventory levels will remain. The “optimal” solution will vary by firm, and over time as the inherent pattern of demand changes. These dynamics and tradeoffs are discussed in depth in [49,100].

In a typical supply chain, there are multiple stages connected in series, for example, in the automotive industry: dealers, car manufacturers/assemblers, machine tool producers, parts manufacturers, raw material suppliers (with potentially several stock management stages in some of these main categories). The upstream stages suffer greater amplification than the downstream stages. In the main, this occurs because each stage uses as its demand signal the orders from the prior stage, which are amplified by that stage’s stock management policies as discussed above. Other reasons for increased upstream amplification include [2]: (1) in determining“expected demand”, each stage tends to extrapolate trends in orders from the prior stage; (2) order batching; (3) price fluctuations (in response to inventory levels); and (4) shortage gaming (ordering more than you really need to get a higher share of the rationed goods from the supplier; see [52] for an example in the aircraft industry). System dynamics analyzes have identified a number of structural changes which can improve supply chain and stock management performance: (1) reduce delays; (2) reduce inventory; and (3) share information (for example, if upstream stages are aware of the downstream end user customer demand pattern, they can use that information rather than the amplified orders from the prior stage as the basis for their decisions, and at least partially avoid amplification [20].

Applications of system dynamics to supply chain management, and production management, remain an important area of research and applications. Akkermans and Daellaert [1], in an article entitled “The Rediscovery of Industrial Dynamics: The Contribution of System Dynamics to Supply Chain Management in a Dynamic and Fragmented World”, provide an excellent survey of supply chain management and system dynamics potential role in moving that field forward. Additional work in this area includes Morecroft’s original analysis of the dynamics created by MRP systems ([64]); Gonçalves doctoral dissertation [32], some of which is summarized in [33,34]; Anderson and Fine [3] on capital equipment supply cycles, and Zahn et al. [114] on flexible assembly systems. Each of these discusses variations on the basic stock/production/supply chain management systems, and provides references for further research.

In addition to inventories, firms need to manage other stocks and resources, including raw materials, employees, capital equipment, and so on; the stock management structure described above for inventory applies to these other stocks as well. The management of stocks and resources is central to dynamic and strategy problems in many industries. First, the management of one stock often influences the ability to manage other stocks (for example, capital equipment and employees determine production). Not only does this interdependency create constraints, the additional negative feedback control in managing resources is another source of cyclical behavior (see Chap. 19 in [100], and Chap. 5 of [69]). Second, in addition to the stocks of resources, production is affected by the productivity of those resources. Dynamic drivers of productivity, such as experience and fatigue, are discussed in the next section.

While the negative control feedbacks described above are central to the observed cyclical behavior of supply chains and resource-based firms, an additional negative feedback through the market adds a further source of instability. This dynamic is perhaps clearest in commodity-based industries, which have also been extensively modeled by system dynamicists as first summarized by Meadows [58]. As illustrated in Fig. 7, these models integrate the supply chain with the dynamics created by supply and demand – in a typical commodity system, there are three major negative feedback loops: two supply feedbacks (one through production, often representing the resource labor, and one through the resource capacity), and one demand feedback (for example, an increase in inventory causes prices to fall, which increases demand and leads to a decrease in inventory from what it otherwise would be). Commodity industries typically exhibit behaviors that include cycles of two periodicities, one determined primarily by the shorter production feedback loop and another longer cycle driven by the capacity loop (see Chap. 20 in [100] for both detailed equations and for examples of



Business Policy and Strategy, System Dynamics Applications to, Figure 7
Commodity industry dynamics showing three controlling feedbacks (adopted from [100])

the structure applied to the livestock and paper industries). The demand feedback loop, however, can play a role in the dynamics as well – if the demand feedback is strong and with a short delay, then demand corrections occur before the supply feedbacks operate and system stability is improved; however, if the magnitude of the delay in the demand loop is similar to the magnitude of the delays in either of the supply loops, the intensity of the corresponding cycle is increased as two negative feedback loops are both independently acting to "solve" the inventory problem. In addition, commodity industries, where they involve a depletable resource such as oil, can experience long-term resource depletion dynamics [101].

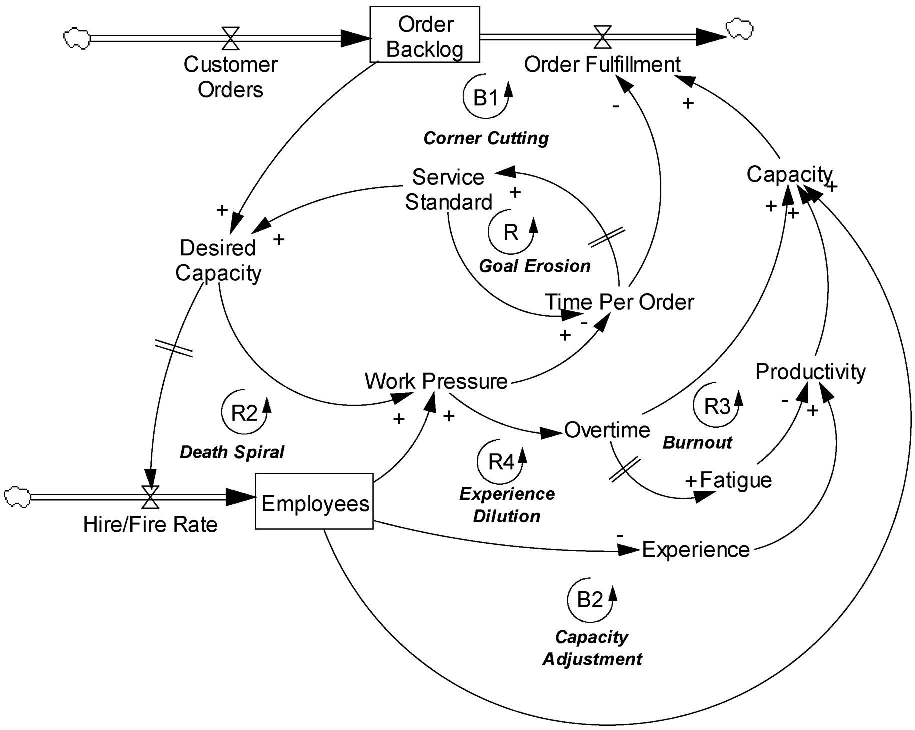
In conclusion, manufacturing and supply chain dynamics are central to many of the behaviors observed in businesses (see Chap. 20 in [100] for real world examples of these cycles). The supply chain, stock management, resource management, and commodity structures discussed above are therefore important components of many system dynamics models developed to support business policy and strategy. In some cases, the firm can change policies to reduce the severity of these cycles; in other cases, especially where the cycles are driven primarily by industry dynamics, the individual firm can use the enhanced understanding and forecasting of these cycles for more strategic decisions such as new product introduction and capacity planning (as in the commercial aircraft market case illustrated in Fig. 1 and discussed in Sect. Applications and Case Examples).

Productivity and Eroding Performance: Service Industry Dynamics

Service-based firms (e. g. professional services, transportation, catalog and online shopping, etc.), and the service arms of manufacturing-based organizations, have a somewhat different set of structural dynamics. Firms in these industries have a number of characteristics that make them more difficult to manage than more manufacturing intensive industries: (1) their product is difficult if not impossible to inventory, and so something else must buffer changes in demand; (2) they are particularly dependent on the performance of people (although the productivity of resources is also important to manufacturing-based businesses as well); and (3) the performance of the system can be harder to detect, and so they are much more subject to a gradual erosion in performance and goals. "The major recurring problems observed in service industry – erosion of service quality, high turnover, and low profitability – can be explained by the organization's response to changes in work pressure." (see p. 28 of [75]).

One of the primary distinguishing features of service-based firms is that their end product is people-dependent and cannot be inventoried. While there may be inventories of products that support delivery of the service, that delivery must be performed based on current resources. As a result, work or order backlogs are the stock that buffers demand from"production". A simplified example is shown in Fig. 8. Customer orders (demand) fill an order backlog, which is depleted by order fulfillment. Order fulfillment is based on the firm's service"capacity" and the amount of time spent per order – for the same capacity, order fulfillment will be greater if less time is spent per order (although service quality may suffer). Capacity is dependent upon people, overtime, and productivity, as discussed further below. Employees are increased or decreased based on desired capacity, which in turn depends on order backlog relative to the firm's service standard (time per order). Work pressure depends on desired capacity relative to the current stock of employees – to the extent the number of employees does not increase as needed, work pressure builds which can result in increases in effective capacity via overtime, or reductions in time spent per order. Time spent per order depends on the firm's service standard for time spent per order, modified by work pressure – if work pressure is high, time spent can be reduced. The service standard often responds to actual performance.

Another important characteristic of service supply firms shown in Fig. 8 is that their capacity is particularly dependent on the performance of people, both in numbers and in productivity. While productivity is also a factor in manufacturing systems, the sensitivity of performance to people factors is generally less than in service-based firms. Therefore, models of such firms generally represent in some detail the factors that drive productivity, including: (1) skill and experience, often using an aging chain or "rookie-pro" structure [100,113]; (2) fatigue from sustained amounts of overtime; and (3) work intensity increasing productivity, but with "haste-makes-waste" impacts on errors and rework (not shown in Fig. 8). In these situations, when demand is growing there are considerable short-term forces which reduce productivity and cause a deterioration in service quality: adding people reduces productivity because of "experience dilution"; working overtime increases fatigue and reduces productivity; pressures to work more intensely increase errors and cause additional work. As shown in Fig. 8, these productivity effects form reinforcing feedback loops which can drive down a service system's performance: an increase in work backlog and desired capacity causes the firm to hire more people; experience levels and productivity decline as a result, thereby reducing order fulfillment below



Business Policy and Strategy, System Dynamics Applications to, Figure 8
Drivers of service business dynamics

what it otherwise would have been; order backlog does not fall as much as expected, necessitating additional capacity, further hires, and decreased experience; this completes the "experience dilution" R4 loop. The "burnout" loop through overtime and fatigue is similarly a reinforcing loop (R3).

Oliva [75] shows that how management responds to these work pressure problems can determine the long-term success or failure of a service-based organization, largely as a result of the third characteristic of such systems: performance of the system can be harder to detect, and so they are much more subject to a gradual erosion in performance and goals. Oliva demonstrates that if the firm reduces its service standards (goals) in response to deteriorating performance (loop R1 goal erosion), a death spiral can ensue in which the declining goals cause the firm to add fewer people, which locks in a situation of excessive work pressure and further declining performance (thereby completing the "death spiral" loop R2). Unless there is a cyclical downturn in demand which alleviates the pressure, a firm's service performance will gradually erode until competition captures the market. He further discusses solutions to these problems, including buffers and faster response. Oliva's work, together with applications noted below, suggest that a service company should hire steadily rather than in spurts to avoid problems of inexperience, should hire enough workers to avoid overwork and a drift to low standards, and (in the case of equipment service) should give preventive maintenance high priority to avoid a spiral of equipment failures.

The resultant financial pressures engendered by the dynamics described above often drive service organizations to investments in process improvement and other cost containment initiatives to seek efficiency gains. Such investments, while offering perhaps the only long-term solution to remaining competitive, cause short-term workloads that further increase the demands on service personnel. This is demonstrated in the work of Repenning and Kaufman [83], and Repenning and Sterman [84,85].

Akkermans and Vos [2], and Anderson et al. [4] have studied the extent to which service industries have multi-stage supply chains similar to manufacturing industries, albeit with backlogs rather than inventories. Akkermans and Vos demonstrate that "inventory" cycles in service chains manifest themselves in terms of order backlog and workload cycles, and that while some of the causes of amplification existent in product supply chains apply to service supply chains (demand signaling and pricing), others, particularly those related to inventory management, do not (order batching, shortage gaming). They find that the real drivers of amplification in service supply chains come from the interactions of workloads, process quality, and rework. Because of delays in hiring and firing, capacity is slow to respond to changes, and is likely to exacerbate cycles. Anderson et al. [4] find that the bullwhip effect may or may not occur in service supply chains, depending on the policies used to manage each stage. However, when it does occur, they find that the systemic improvements that can often be achieved in physical supply chains by locally applied policies (e. g., reducing delay times and sharing information) do not have as many parallels in service chains. Instead service supply chains are characterized by numerous tradeoffs between improving local performance and improving system performance.

The modeling of service delivery has also had a long history in system dynamics, though the number of published works is more modest than in other areas. Much of the early work was in the area of health care and education [46], Later works of note include models of People Express Airlines [98], Hanover Insurance claims processing [96]. NatWest Bank lending [74], and DuPont chemical plant equipment maintenance [12]. Homer [40] presents a case application for a major producer of equipment for semiconductor manufacturing that demonstrates many of the structures and policy issues enumerated above. These works incorporate the basic dynamic theory discussed above and illustrated in Fig. 8, and add another set of important structural theories to the building blocks for business strategy applications (note that the modeling of various effects on productivity is much more extensive in the area of project modeling, as discussed in [55]).

Life Cycles of Products and Diffusion

Another important pattern of behavior characteristic of many firms (or subsets of firms) is that of a life cycle (for the flow) and S-shaped pattern for the stock, as illustrated in Fig. 9: a gradual increase from a low level up to a peak, followed by a gradual decline either to zero or to some replacement level (sometimes referred to as "boom and bust" behavior). Sterman [100] and Oliva et al. [76] provide some real world examples of this behavior. The example shown is common for the sales of new products: the flow represents people becoming customers, and the stock, customers.

The structure which creates this "boom and bust" dynamics is shown in Fig. 10. In the marketing literature this structure is referred to as the "Bass Diffusion Model" after its original proponent [8]. The structure consists of three feedback loops: a reinforcing "word of mouth" loop that dominates behavior in the first half of customer sales growth; a balancing "market saturation" loop that constrains and eventually shuts down growth as the number of potential customers falls to zero; and another balancing loop "advertising saturation", which represents other means of stimulating awareness, such as advertising, direct sales efforts, and media reports. These other channels are usually assumed to be proportional to the size of the pool of potential customers, and therefore initially stimulate the flow of "becoming customers" but then decline over time as the pool is depleted.

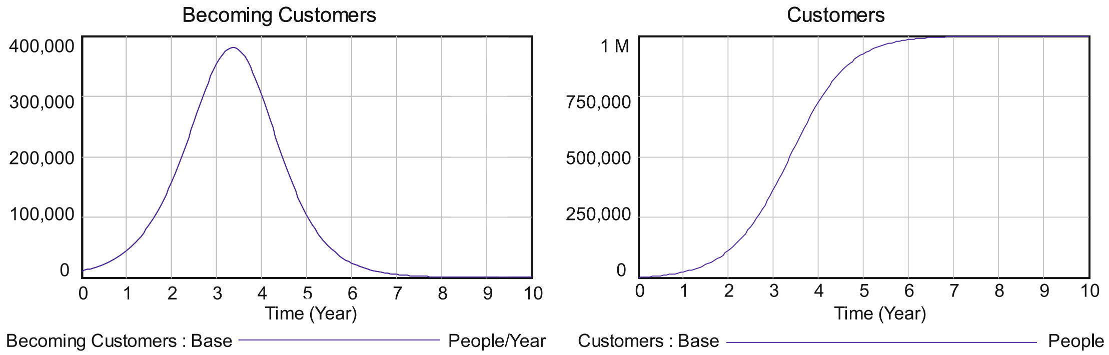
The dynamics of this structure, extensions to it (for example, loss of customers, competition, repeat sales), and policy implications are discussed in depth in Chap. 9 in [100] and Chap. 6 in [69]. This structure forms the basis of many system dynamics models that represent product sales, customer development, and the diffusion of innovations. Published examples include the work of Milling [60,61] and Maier [57] on the management of innovation diffusions and Oliva et al. [76] on boom and bust in e-commerce. Milling discusses the ► Diffusion of Innovations, System Dynamics Analysis of the in more depth.

Growth Dynamics

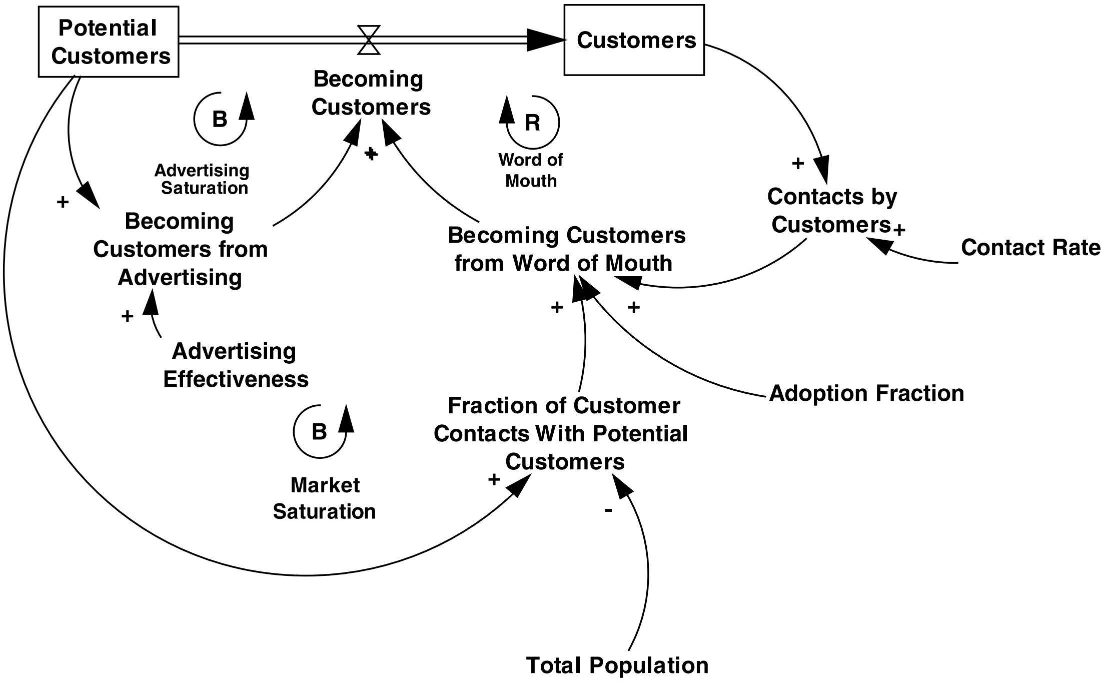
Growth is fundamentally a dynamic process, and therefore it is no surprise that since its early days system dynamicists have shown an interest in the dynamics of corporate growth. Forrester [27,28], Packer [77], Lyneis [48,49], Morecroft [67], Morecroft and Lane [70] and others studied corporate growth in the field's early years. More recently, the People Express [98] and B&B ("Boom and Bust") flight simulators [78] illustrate the field's interest in growth dynamics.

In his 1964 article, Forrester identified the range of possible growth patterns (see Fig. 11): smooth, steady growth; growth with repeated setbacks; stagnation; and decline. Examples of these patterns can be found in many real world industries, as illustrated for the computer industry in Fig. 11 (see also [50] for examples). In his classic article "Market Growth as Influenced by Capital Investment", Forrester detailed the three types of feedback loops which can create the range of possible growth patterns. These are illustrated in Fig. 11 (the equations for this model are provided in the original Forrester article; the model is also presented and discussed and analyzed in detail in Chap. 15 in [100], and Chap. 7 in [69]).

On the left in Fig. 12 is the reinforcing "salesforce expansion" loop: the salesforce generates sales, a portion of those sales are allocated to future marketing budgets,

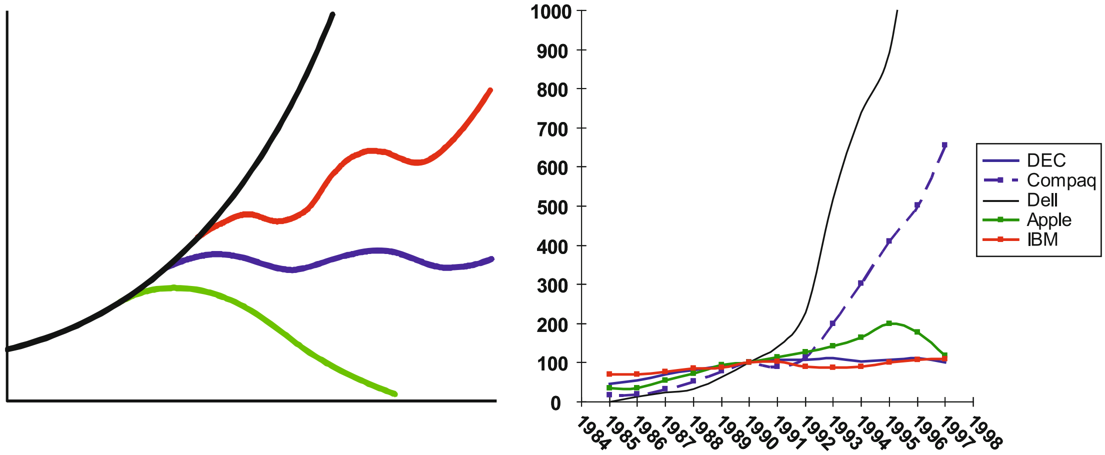


Business Policy and Strategy, System Dynamics Applications to, Figure 9
Life cycle behavior mode ("Boom and Bust")

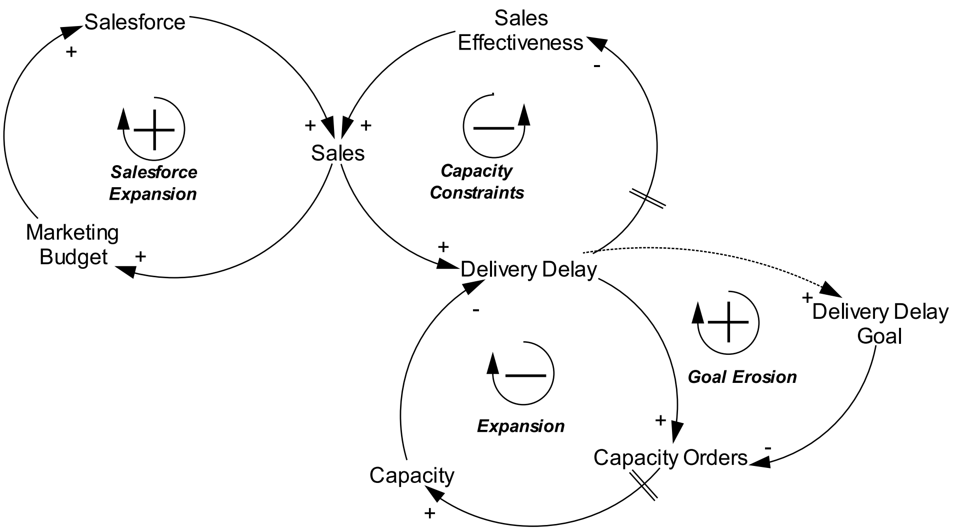


Business Policy and Strategy, System Dynamics Applications to, Figure 10
Basic structure generating boom and bust dynamics

which allows an increase in the size of the salesforce and a further increase in sales. The salesforce expansion loop in isolation can create smooth growth forever (until the market is saturated). However, assuming a fixed capacity, the balancing "capacity constraints" loop activates: if sales exceed capacity, delivery delay increases such that, after a delay, sales effectiveness falls and sales decline. The goal of the loop is to equate sales and capacity, and the two loops together can produce growth followed by stagnation (with fluctuations caused by delays in the balancing loop). In response to increasing delivery delay, however, firms often increase capacity ("capacity expansion" loop): when delivery delay exceeds the firm's delivery delay goal, capacity orders increase, which after a delay increases capacity and thereby reduces delivery delay; the goal of this loop is to equate delivery delay to the firm's delivery delay goal. However, once delivery delay is reduced, sales effectiveness and sales increase, thereby stimulating additional salesforce expansion, such that the growth with setbacks pattern of behavior can result. The final loop shown in Fig. 12 is the reinforcing "goal erosion" loop: if the firm's delivery delay goal responds to the actual delivery delay



Business Policy and Strategy, System Dynamics Applications to, Figure 11
Stylized patterns of growth and examples from the computer hardware industry



Business Policy and Strategy, System Dynamics Applications to, Figure 12
Feedback loops creating observed patterns of growth (adapted from Forrester 1968)

performance, a downward spiral can ensue – the goal increases, less expansion occurs than had before, capacity is less than needed, delivery delay increases, the goal is further increased, and so on (this loop is similar to the service standard goal erosion loop discussed above). The goal erosion loop can create the decline dynamics illustrated in Fig. 11 (although in actual practice the decline would likely occur over a much more extended period than shown).

In actual practice, there are numerous positive feedback loops through resources that might stimulate growth. These loops are listed below, and many are discussed and diagrammed in Chap. 10 in [100]. In each of these loops, an increase in sales causes management actions and/or investments that further increase the resource and sales:

- Sales channels – sales capability (which might include salesforce as discussed above, or retail stores), advertising, word-of-mouth contagion (as in the diffusion model), media hype (sales create media exposure which attracts potential customers and more sales)
- Price – Product attractiveness channels (operationalizing the link between resources and costs in Fig. 2 – spreading of fixed costs over more units, thereby lowering unit costs; learning curves; economies of scale;

economies of scope; investments in process improvements)

- Market channels which increase the pool of potential customers – network effects (the more people using cell phones the greater their value), development of complementary goods (software applications for computers)
- Product investment channels – product improvement, new products
- Market power channels – over suppliers, over labor, over customers, cost of capital.

With all these positive feedback loops, how can anyone fail? In fact, there are also numerous constraints to growth, including: depletion of the pool of potential customers as discussed in the last section; growth of competition; delays in acquiring production capacity and/or service capacity; limits to financial capital (which can increase delays or limit acquiring productive assets); and increases in organizational size, complexity and administrative overheads (which might make the resources – costs loop in Fig. 2 revert to a positive connection, thereby constraining growth). Structures for representing these constraints are provided in the earlier references to this section, especially [49,69,100], and form the building blocks for many of the practical applications of system dynamics to business growth strategy.

As system dynamicists have long recognized, managing growth is one of the more challenging management tasks. It entails fostering the positive, reinforcing feedback loops while simultaneously relaxing the constraining, negative feedback loops. While it is difficult to generalize without sounding platitudinous, a number of important lessons have emerged from studies of growth. Lyneis [50] discusses these in more detail, and references other work:

Lesson 1 You won't achieve what you don't try for (if a firm is timid in its growth objectives, it will be timid in the acquisition of resources – balancing loops through for example delivery delay will then drive sales growth to the firm's resource growth). Corollary 1: Don't mistake forecasts for reality (a firm may be continually surprised by how accurate their sales forecasts are, because if resources are based on these forecasts, the balancing loops will drive sales to those resources). Corollary 2: Provide sufficient buffers and contingencies (these help minimize the risks that the balancing loops will become dominant).

Lesson 2 Avoid the temptation to reduce objectives in the face of performance problems (the "goal erosion" loop in Figs. 8 and 12).

Lesson 3 In a world of limited resources, something must limit growth. Proactively managing these limits is a key factor affecting performance. For example, if financial constraints are limiting expansion, with resultant delivery constraints on sales, why not increase prices to limit sales, and use the extra cash to finance expansion?

Lesson 4 Make an effort to account for delays, especially in market response (for example, improvements in service will take a while to manifest themselves in improved sales, so avoid the temptation to cut back productive resources that will later be needed).

Lesson 5 Account for short-term productivity losses such as fatigue and experience dilution in resource expansion decisions (in the short-term, you may be getting less capacity than you think).

Lesson 6 Account for likely competitor responses in taking actions (it's easy for competitors to follow price changes, and trigger a price war; improvements in other components of product attractiveness are harder to detect and replicate).

Applications and Case Examples

The process and structural developments discussed in the last sections have formed the basis of numerous applications of system dynamics in support of business strategy. In turn, these applications have provided the practice field through which the process has been refined, and the structural models, insights, and tools validated. While most of the published "real world" applications of system dynamics do not provide details of the models, they do nevertheless provide diagrams which show the nature of the model, representative results and policy conclusions, and the role the models played in business strategy formulation.

Space limitations preclude covering specific applications in depth. Therefore, I have chosen to reference applications in a number of different areas so readers can find references to the literature in their particular area of interest. Before getting to that, however, there are a couple of general references worthy of note: Roberts [91] covers many of the early published applications of system dynamics, with sections on manufacturing, marketing, research and development, and management and financial control. Coyle [14,15,16] touches on many of the applications initiated by his team at the University of Bradford, including work in the defense, natural resources, and utility industries. Richardson [86] provides an edited collection of academic journal articles containing some of the best work in system dynamics for business (and public) policy from its early years to the 1990s. Beyond these general compendiums, business strategy applications can perhaps best be classified by the industry of interest.

First, there have been a number of industry-wide models. The purpose of these models is typically to understand the drivers of change in the industry, and to forecast demand for use in other planning models. These include:

- Aircraft market as illustrated in Fig. 1 above [47,53]
- Health care market [42,103]
- Oil market [69,71,101]
- Shipping market [82].

Second, in addition to these industry-wide models, multiple applications to specific firms (which might also include some industry and/or competitive modeling), have been done in the following industry sectors:

- Utilities/Regulation – In the electric industry, work by Ford [24], Lyneis [51], Bunn and Larson [10,45] Ford covers some of this work elsewhere in ▶ System Dynamics Models of Environment, Energy and Climate Change.
- Telecoms [35]
- Financial Services – work for MasterCard [52] and in the insurance industry [7,13,96,102].

Finally, applications to specific types of firms, particularly small and medium size enterprises [9].

For particular examples of where system dynamics has had an impact in changing or forming business strategy, the MasterCard application described by Lyneis [52] and the General Motors OnStar application described by Barabba et al. [6] are noteworthy. These applications provide some detail about the model structure and describe how the modeling process changed management intuition and thereby led to significant shifts in business strategy. The MasterCard model represents growth and competitive dynamics in some detail, and is used to illustrate the multi-stage development process detailed in Sect. Using System Dynamics Models in Policy and Strategy Formulation and Implementation above. The OnStar example describes the process of modeling an industry that does not yet exist. The model itself builds from the diffusion model discussed above, with significant elaboration of potential customers, provision of service, alliances, dealers, and financial performance. The paper details the significant role that the system dynamics model played in reshaping GM's strategy.

Future Directions

System dynamics has made significant theoretical and practical contributions to business strategy. These contributions fall into two general categories: first, the process through which models are developed, working with management teams to enhance model validity and implementation; and second, understanding of the business structures and policies which cause observed problem behavior within firms and industries. Nevertheless, the impact of system dynamics on business strategy has been modest – relatively few firms use system dynamics. In my view, several factors contribute to this slow uptake. These are listed below, with possible actions that could be taken to alleviate.

1. Knowledge of system dynamics and its potential in strategy formulation is limited. Senior executives of firm's are unaware of system dynamics, or its potential [106]. In part this is because system dynamics is taught at relatively few business schools. Over time, this problem will be solved, but it will take a long time. But more importantly, researchers and practitioners of system dynamics rarely publish their work in publications which are widely read by senior management, such as the Harvard Business Review. This is a problem which can be addressed in the near future if researchers and practitioners made the effort to communicate with this market.
2. System dynamics is not well connected with traditional strategy research and practice. As noted earlier, system dynamics and traditional strategy developed largely independently. While the potential synergies between the two are significant, until recently few researchers and practitioners have made the effort to cross disciplines. More effort and publication are needed to demonstrate areas of synergy and get system dynamics into the mainstream of business strategy research and ultimately practice. Warren [108] makes a start on this.
3. System dynamics is hard. Building system dynamics models, calibrating them to data, and analyzing their behavior to improve business strategy requires significant skill and experience. This has traditionally been developed via an apprenticeship program, either in university or consulting firms. Much more can be done to hasten this process. First, tried and true model structures that can be used as building blocks for models must be better documented, and gaps closed. While the theoretical basis for business dynamics described above is a good start, and reflects structures which all practicing system dynamicists should know, the underlying models and building blocks are widely dispersed and difficult to access. For example, the models in [100] are contained within a 900 page introductory textbook; the models in [49] are in old software and the book is

out of print. In both cases, the models are only starting points and much unpublished work has occurred that expands these introductory models to make them more relevant and directly applicable to real business strategy problems. Efforts could and should be made to expand the library of business strategy building blocks. In addition, the process of building models and working with managers needs to be better documented and disseminated, both in formal courses and in published works, so as to facilitate the apprenticeship learning process.

Bibliography

1. Akkermans HA, Daellaert N (2005) The Rediscovery of Industrial Dynamics: The Contribution of System Dynamics to Supply Chain Management in a Dynamic and Fragmented World. Syst Dyn Rev 21(3):173–186
2. Akkermans HA, Vos B (2003) Amplification in service supply chains: an exploratory case study from the telecom industry. Prod Oper Manag 12(2):204–223
3. Anderson E, Fine C (1999) Business Cycles and Productivity in Capital Equipment Supply Chains. In: Tayur et al(eds) Quantitative Models for Supply Chain Management. Kluwer Academic Publishers, Norwell
4. Anderson E, Morrice D, Lundeen G (2005) The 'physics' of capacity and backlog management in service and custom manufacturing supply chains. Syst Dyn Rev 21(3):187–216
5. Backus G, Schwein MT, Johnson ST, Walker RJ (2001) Comparing expectations to actual events: the post mortem of a Y2K analysis. Syst Dyn Rev 17(3):217–235
6. Barabba V, Huber C, Cooke F, Pudar N, Smith J, Paich M (2002) A Multimethod Approach for Creating New Business Models: The General Motors OnStar Project. Interfaces 32(1):20–34
7. Barlas Y, Cırak K, Duman E (2000) Dynamic simulation for strategic insurance management. Syst Dyn Rev 16(1):43–58
8. Bass FM (1969) New product growth model for consumer durables. Manag Sci 15:215–227
9. Bianchi C (2002) Editorial to Special Issue on Systems Thinking and System Dynamics in Small-Medium Enterprises. Syst Dyn Rev 18(3):311–314
10. Bunn DW, Larsen ER (eds) (1997) Systems Modeling for Energy Policy. Wiley, Chichester
11. Campbell D (2001) The long and winding (and frequently bumpy) road to successful client engagement: one team's journey. Syst Dyn Rev 17(3):195–215
12. Carroll JS, Sterman JD, Marcus AA (1998) Playing the maintenance game: How mental models drive organizational decisions. In: Halpern JJ, Stern RN (eds) Nonrational Elements of Organizational Decision Making. Cornell University Press, Ithaca
13. Doman A, Glucksman M, Mass N, Sasportes M (1995) The dynamics of managing a life insurance Company. Syst Dyn Rev 11(3):219–232
14. Coyle RG (1996) System Dynamics Modelling: A Practical Approach. Chapman and Hall, London
15. Coyle RG (1997) System Dynamics at Bradford University: A Silver Jubilee Review. Syst Dyn Rev 13(4):311–321
16. Coyle RG (1998) The Practice of System Dynamics: Milestones, Lessons and Ideas From 30 Years Experience. Syst Dyn Rev 14(4):343–365
17. Coyle RG (2000) Qualitative and Quantitative Modeling in System Dynamics: Some Research Questions. Syst Dyn Rev 16(3):225–244
18. Coyle RG (2001) Rejoinder to Homer and Oliva. Syst Dyn Rev 17(4):357–363
19. Coyle RG, Morecroft JDW (1999) System Dynamics for Policy, Strategy and Management Education. J Oper Res Soc 50(4)
20. Croson R, Donohue K (2005) Upstream versus downstream information and its impact on the bullwhip effect. Syst Dyn Rev 21(3):187–216
21. Delauzun F, Mollona E (1999) Introducing system dynamics to the BBC World Service: an insider perspective. J Oper Res Soc 50(4):364–371
22. Doman A, Glucksman M, Mass N, Sasportes M (1995) The dynamics of managing a life insurance company. Syst Dyn Rev 11(3):219–232
23. Dyson RG, Bryant J, Morecroft J, O'Brien F (2007) The Strategic Development Process. In: O'Brien FA, Dyson RG (eds) Supporting Strategy. Wiley, Chichester
24. Ford AJ (1997) System Dynamics and the Electric Power Industry. Syst Dyn Rev 13(1):57–85
25. Forrester JW (1958) Industrial dynamics: a major breakthrough for decision makers. Harv Bus Rev 36(4):37–66
26. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge (now available from Pegasus Communications, Waltham)
27. Forrester JW (1964) Common Foundations Underlying Engineering and Management. IEEE Spectr 1(9):6–77
28. Forrester JW (1968) Market growth as influenced by capital investment. Industrial Management Review 9(2):83–105. Reprinted in Forrester JW (1975) Collected Papers of Jay W Forrester. Pegasus Communications, Waltham
29. Forrester JW (1969) Urban Dynamics. Pegasus Communications, Waltham
30. Foss NJ (ed) (1997) Resources, Firms and Strategies. Oxford University Press, Oxford
31. Gary MS (2005) Implementation Strategy and Performance Outcomes in Related Diversification. Strateg Manag J 26:643–664
32. Gonçalves PM (2003) Demand Bubbles and Phantom Orders in Supply Chains. Unpublished Dissertation Sloan School of Management. MIT, Cambridge
33. Gonçalves PM (2006) The Impact of Customer Response on Inventory and Utilization Policies. J Bus Logist 27(2):103–128
34. Gonçalves P, Hines J, Sterman J (2005) The Impact of Endogenous Demand on Push-Pull Production Systems. Syst Dyn Rev 21(3):187–216
35. Graham AK, Walker RJ (1998) Strategy Modeling for Top Management: Going Beyond Modeling Orthodoxy at Bell Canada. Proceedings of the 1998 International System Dynamics Conference, Quebec
36. Grant RM (2005) Contemporary Strategy Analysis, 5th edn. Blackwell, Cambridge
37. Hines JH, Johnson DW (1994) Launching System Dynamics. Proceedings of the 1994 International System Dynamics Conference, Business Decision-Making, Stirling

38. Homer JB (1996) Why We Iterate: Scientific Modeling In Theory and Practice. Syst Dyn Rev 12(1):1–19
39. Homer JB (1997) Structure, Data, and Compelling Conclusions: Notes from the Field. Syst Dyn Rev 13(4):293–309
40. Homer JB (1999) Macro- and Micro-Modeling of Field Service Dynamics. Syst Dyn Rev 15(2):139–162
41. Homer J, Oliva R (2001) Maps and Models in System Dynamics: A Response to Coyle. Syst Dyn Rev 17(4):347–355
42. Homer J, Hirsch G, Minniti M, Pierson M (2004) Models for Collaboration: How System Dynamics Helped a Community Organize Cost-Effective Care For Chronic Illness. Syst Dyn Rev 20(3):199–222
43. Kim DH, Lannon C (1997) Applying Systems Archetypes. Pegasus Communications Inc, Waltham
44. Kunc M, Morecroft J (2007) System Dynamics Modelling for Strategic Development. In: O'Brien FA, Dyson RG (eds) Supporting Strategy. Wiley, Chichester
45. Larsen ER, Bunn DW (1999) Deregulation in electricity: understanding strategic and regulatory risks. J Oper Res Soc 50(4):337–344
46. Levin G, Roberts EB, Hirsch GB (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
47. Liehr M, Großler A, Klein M, Milling PM (2001) Cycles in the sky: understanding and managing business cycles in the airline market. Syst Dyn Rev 17(4):311–332
48. Lyneis JM (1975) Designing Financial Policies to Deal With Limited Financial Resources. Financ Manag 4(1)
49. Lyneis JM (1980) Corporate Planning and Policy Design: A System Dynamics Approach. M.I.T. Press, Cambridge
50. Lyneis JM (1998) Learning to Manage Growth: Lessons From a Management Flight Simulator, Proceedings of the 1998 International System Dynamics Conference (Plenary Session), Quebec City, 1998
51. Lyneis JM (1997) Preparing for a Competitive Environment: Developing Strategies for America's Electric Utilities. In: Bunn DW, Larsen ER (eds) Systems Modeling for Energy Policy. Wiley, Chichester
52. Lyneis JM (1999) System dynamics for business strategy: a phased approach. Syst Dyn Rev 15(1):37–70
53. Lyneis JM (2000) System Dynamics for Market Forecasting and Structural Analysis. Syst Dyn Rev 16(1):3–25
54. Lyneis JM, Cooper KG, Els SA (2001) Strategic Management of Complex Projects: A Case Study Using System Dynamics. Syst Dyn Rev 17(3):237–260
55. Lyneis JM, Ford DN (2007) System Dynamics Applied to Project Management: A Survey, Assessment, and Directions for Future Research. Syst Dyn Rev 23(2/3):157–189
56. Magee J (1958) Production Planning and Inventory Control. McGraw-Hill, London
57. Maier FH (1998) New product diffusion models in innovation management – a system dynamics perspective. Syst Dyn Rev 14(4):285–308
58. Meadows DL (1970) Dynamics of commodity production cycles. Pegasus Communications, Waltham
59. Merten PP, Loffler R, Wiedmann KP (1987) Portfolio Simulation: A Tool to Support Strategic Management. Syst Dyn Rev 3(2):81–101
60. Milling PM (1996) Modeling innovation processes for decision support and management Simulation. Syst Dyn Rev 12(3):211–23
61. Milling PM (2002) Understanding and managing innovation processes. Syst Dyn Rev 18(1):73–86
62. Milling PM (2007) A Brief History of System Dynamics in Continental Europe. Syst Dyn Rev 23(2/3):205–214
63. Montgomery CA (ed) (1995) Resource-Based and Evolutionary Theories of the Firm. Kluwer, Boston
64. Morecroft JDW (1983) A Systems Perspective on Material Requirements Planning. Decis Sci 14(1):1–18
65. Morecroft JDW (1984) Strategy Support Models. Strateg Manag J 5:215–229
66. Morecroft JDW (1985) The Feedback View of Business Policy and Strategy. Syst Dyn Rev 1(1):4–19
67. Morecroft JDW (1986) The Dynamics of a Fledgling High-Technology Growth Market. Syst Dyn Rev 2(1):36–61
68. Morecroft JDW (1999) Management attitudes, learning and scale in successful diversification: a dynamic and behavioural resource system view. J Oper Res Soc 50(4):315–336
69. Morecroft JDW (2007) Strategic Modelling and Business Dynamics: A Feedback Systems View. Wiley, Chichester
70. Morecroft JDW, Lane DC, Viita PS (1991) Modeling growth strategy in a biotechnology startup firm. Syst Dyn Rev 7(2):93–116
71. Morecroft JDW, van der Heijden KAJM (1994) Modeling the Oil Producers: Capturing Oil Industry Knowledge in a Behavioral Simulation Model. In: Morecroft JDW, Sterman JD (eds) Modeling for Learning Organizations. Productivity Press, Portland
72. Morecroft JDW, Wolstenholme E (2007) System Dynamics in the UK: A Journey from Stirling to Oxford and Beyond. Syst Dyn Rev 23(2/3):205–214
73. O'Brien FA, Dyson RG (eds) (2007) Supporting Strategy. Wiley, Chichester
74. Oliva R (1996) A Dynamic Theory of Service Delivery: Implications for Managing Service Quality. Ph.D. Thesis, Sloan School of Management, Massachusetts Institute of Technology
75. Oliva R (2001) Tradeoffs in responses to work pressure in the service industry. Calif Manag Rev 43(4):26–43
76. Oliva R, Sterman JD, Giese M (2003) Limits to growth in the new economy: exploring the 'get big fast' strategy in e-commerce. Syst Dyn Rev 19(2):83–117
77. Packer DW (1964) Resource Acquisition in Corporate Growth. M.I.T. Press, Cambridge
78. Paich M, Sterman JD (1993) Boom, Bust, and Failures to Learn in Experimental Markets. Manag Sci 39(12)
79. Porter ME (1980) Competitive Strategy. The Free Press, New York
80. Porter ME (1985) Competitive Advantage. The Free Press, New York
81. Prahalad CK, Hamel G (1990) The Core Competence of the Corporation. Harv Bus Rev (May-June):71–91
82. Randers J, Göluke U (2007) forecasting turning points in shipping freight rates: lessons from 30 years of practical effort. Syst Dyn Rev 23(2/3):253–284
83. Repenning N, Kofmann F (1997) Unanticipated Side Effects of Successful Quality Programs: Exploring a Paradox of Organizational Improvement. Manag Sci 43(4)
84. Repenning NP, Sterman JD (2001) Nobody Ever Gets Credit for Fixing Problems that Never Happened: Creating and Sustaining Process Improvement. Calif Manag Rev 43(4):64–88

85. Repenning NP, Sterman JD (2002) Capability traps and self-confirming attribution errors in the dynamics of process improvement. Adm Sci Q 47:265–295
86. Richardson GP (1996) Modelling for Management: Simulation in Support of Systems Thinking. Dartsmouth Publishing Company, Aldershot
87. Richardson GP, Andersen DF (1995) Teamwork in Group Model Building. Syst Dyn Rev 11(2):113–138
88. Richmond B (1997) The Strategic Forum: Aligning Objectives, Strategy and Process. Syst Dyn Rev 13(2):131–148
89. Risch J, Troyano-Bermúdez L, Sterman J (1995) Designing Corporate Strategy with System Dynamics: A Case Study in the Pulp and Paper Industry. Syst Dyn Rev 11(4):249–274
90. Roberts EB (1977) Strategies for the Effective Implementation of Complex Corporate Models. Interfaces 8(1):26–33
91. Roberts EB (ed) (1978) Managerial Applications of System Dynamics. The MIT Press, Cambridge
92. Roberts EB (2007) Making System Dynamics Useful: A Personal Memoir. Syst Dyn Rev 23(2/3):119–136
93. Rockart SF, Lenox MJ, Lewin AY (2007) Interdependency, Competition, and Industry Dynamics. Manag Sci 53(4):599–615
94. Snabe B, Grossler A (2006) System Dynamics Modelling for Strategy Implementation—Case Study and Issues. Syst Res Behav Sci 23:467–481
95. Senge PM (1990) The Fifth Discipline: The Art and Practice of The Learning Organization. Doubleday, New York
96. Senge PM, Sterman JD (1992) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59(1):137–150
97. Sice P, Mosekilde E, Moscardini A, Lawler K, French I (2000) Using system dynamics to analyse interactions in duopoly competition. Syst Dyn Rev 16(2):113–133
98. Sterman JD (1989) Modeling management behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35:321–339
99. Sterman JD (1989) Strategy Dynamics: the Rise and Fall of People Express. Memorandum D-3939-1 Sloan School of Management, M.I.T
100. Sterman JD (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill, New York
101. Sterman JD, Richardson G, Davidsen P (1988) Modeling the estimation of petroleum resources in the United States. Technol Forecast Soc Chang 33(3):219–249
102. Thompson JP (1999) Consulting approaches with system dynamics: three case studies. Syst Dyn Rev 15(1):71–95
103. Thompson JP (2006) Making sense of US health care system dynamics. Proceedings of the 2006 International System Dynamics Conference, Nijmegen, The Netherlands
104. Vennix JAM (1996) Group Model Building: Facilitating Team Learning Using System Dynamics Field. Wiley, Chichester
105. Warren KD (2002) Competitive Strategy Dynamics. Wiley, Chichester
106. Warren KD (2004) Why Has Feedback Systems Thinking Struggled to Influence Strategy and Policy Formulation? Suggestive Evidence, Explanations and Solutions. Syst Res Behav Sci 21:331–347
107. Warren KD (2005) Improving Strategic Management With the Fundamental Principles of System Dynamics. Syst Dyn Rev 21(4):329–350
108. Warren KD (2007) Strategic Management Dynamics. Wiley, Chichester
109. Weil HB (1980) The Evolution of an Approach for Achieving Implemented Results from System Dynamics Projects. In: Randers J (ed) Elements of the System Dynamics Method. MIT Press, Cambridge
110. Wolstenholme EF (1999) Qualitative vs. Quantitative Modelling: The Evolving Balance. J Oper Res Soc 50(4):422–428
111. Winch GW (1993) Consensus building in the planning process: benefits from a "hard" modeling approach. Syst Dyn Rev 9(3):287–300
112. Winch GW (1999) Dynamic Visioning for Dynamic Environments. J Oper Res Soc 50(4):354–361
113. Winch GW (2001) Management of the "skills inventory" in times of major change. Syst Dyn Rev 17(2):151–159
114. Zahn E, Dillerup R, Schmid U (1998) Strategic evaluation of flexible assembly systems on the basis of hard and soft decision criteria. Syst Dyn Rev 14(4):263–284

Corporate and Municipal Bond Market Microstructure in the U.S.

MICHAEL S. PIWOWAR
Securities Litigation and Consulting Group, Inc.,
Fairfax, USA

Article Outline

Glossary

ABS Automated Bond System. The original automated limit-order market for bonds operated by the NYSE that executed orders according to strict price/time priority. ABS was replaced by the NYSE Bonds Platform in 2007.

Agency trade A bond transaction executed by a broker-dealer on behalf of another party. A broker-dealers is compensated by a commission on an agency trade.

Broker A firm that acts as an intermediary by executing agency trades.

Broker-dealer A firm that engages in both agency trades and principal trades.

Broker's broker A broker-dealer that exclusively executes agency trades of municipal bonds with other broker-dealers. Broker's brokers do not execute principal trades and they do not trade directly with public investors.

Commission A form of compensation that a customer pays a broker-dealer for executing an agency trade. Broker-dealers must explicitly disclose the commission to the customer as a separate item on the customer's trade confirmation.

Dealer A firm that engages in principal trades for its own account.

FINRA Financial Industry Regulatory Authority. The self-regulatory organization (SR0) created in July 2007 from the consolidation of NASD and the member regulation, enforcement and arbitration functions of the NYSE. FINRA rules are approved by the SEC and enforced by themselves.

FIPS Fixed Income Pricing Service. The electronic system operated by the National Association of Securities Dealers (NASD) from 1994 through 2002 to collect and disseminate real-time quotations and hourly trade reports for a subset of high-yield corporate bonds. FIPS was retired in July 2002 with the implementation of TRACE.

Market maker A specific designation made by a regulatory authority for a broker-dealer that holds itself out to trade securities by publishing regular or continuous quotations to buy (bid) or sell (offer). Currently, there are no broker-dealers regulated as market makers in the US corporate or municipal bond markets.

Mark-up and mark-down A form of compensation that a customer pays a broker-dealer for executing a principal trade. Customers pay a mark-up when they buy a bond from a broker-dealer; they pay a mark-down when they sell a bond to a broker-dealer. Unlike commissions, mark-ups and mark-downs do not need to be disclosed on customer trade confirmations.

MSRB Municipal Securities Rulemaking Board. The self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their municipal bond transactions. MSRB rules are approved by the SEC and enforced by FINRA (formerly NASD).

NASD Formerly known as the National Association of Securities Dealers. The self-regulatory organization (SRO) charged with, among other things, primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. In July 2007, NASD and the member regulation, enforcement and arbitration functions of the NYSE consolidated to form FINRA.

NYSE New York Stock Exchange. Operates the NYSE Bonds Platform (formerly ABS) trading system for exchange-listed corporate bonds.

OTC securities Over the-counter securities. Securities that are not traded on an organized exchange.

Principal trade A bond transaction executed by a broker-dealer for its proprietary account. The broker-dealer is compensated by a mark-up or mark-down on a principal trade.

Riskless principal trade A principal trade in which a broker-dealer purchases a bond to satisfy a previously received order to buy, or a broker-dealer sells a bond to satisfy a previously received order to sell. The transaction is riskless to the broker-dealer because the firm does not bear any inventory (price) risk.

RTTRS (or TRS) (Real-Time) Transaction Reporting System. MSRB's municipal bond transaction reporting and dissemination system.

Serial offering A bond issuance in which several different bonds are offered with different, often consecutive, maturities. Municipal bonds are typically issued in serial offerings.

SRO Self-regulatory Organization. A non-governmental industry association that has statutory authority to regulate members through the promulgation and enforcement of rules and regulations governing business practices. The SEC oversees SRO activities and approves SRO rules.

SEC US Securities and Exchange Commission. The primary governmental overseer and regulator of US securities markets, including the corporate and municipal bond markets. Broker-dealers and SROs are overseen by the SEC's Division of Trading and Markets (formerly Division of Market Regulation).

TRACE (formerly NASD TRACE) Transaction Reporting and Compliance Engine. FINRA's corporate bond transaction reporting and dissemination system.

Definition of the Subject

The subject of this article is the microstructure of the US corporate and municipal bond markets. ► Treasury Market, Microstructure of the U.S. provide a complementary discussion of the microstructure of the US Treasury bond market.

Market microstructure is broadly defined as the study of the economics of markets and trading. Market microstructure research covers a wide range of interrelated topics including market structure and design issues (e. g., trading systems and rules); price formation and price discovery; strategic trading behavior; market quality, liquidity, and trading costs (explicit and implicit); information, disclosure, and transparency; and consequences of regulatory policy (intended and unintended).

While much has been written on the microstructure of equity markets since the mid-1980s, the bond markets have only recently started receiving attention from academic researchers. The development of research in both markets can largely be attributed to the availability of quality intraday trade, quote, and/or order data ("tick" data) to empirical researchers.

The seminal theoretical work in market microstructure was conducted contemporaneously with the early equity market microstructure research, and much of the underlying economics is general enough to be appropriate for the bond markets. As a result, the significant contributions of bond market research so far have been almost exclusively empirical in nature. The last study featured in this article by Green, Hollifield, and Schurhoff [23] is a notable exception.

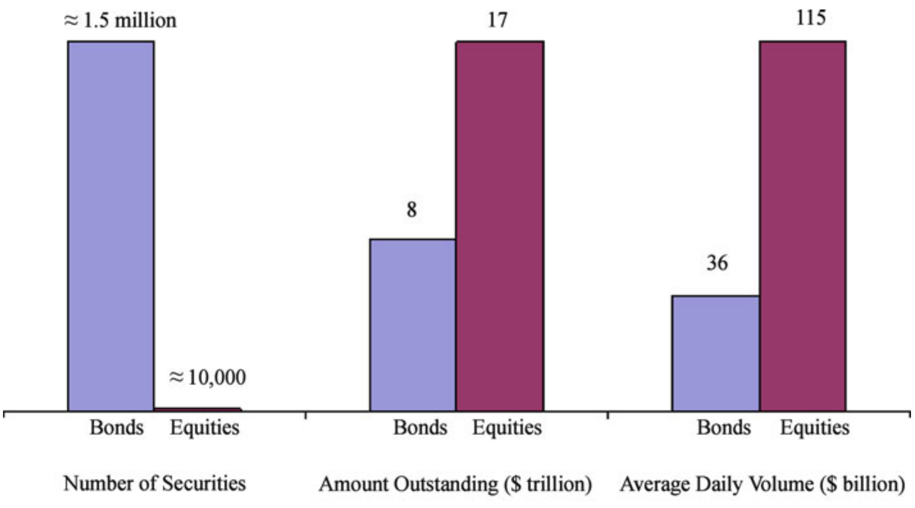
Conversely, the empirical methods developed specifically for the structure and design of equity markets are not well-suited for the bond markets. Accordingly, many of the important contributions of bond market microstructure research stem from not only the results and conclusions, but also from the development of new empirical methods. This article will provide details on some of these methods as well as discuss the important results, conclusions, and policy implications.

But, before moving on to a detailed discussion of bond market microstructure research, an important question needs to be answered. Why should we care about the bond markets? We should care because the bond markets provide an important source of capital for issuers and an important source of securities for investors. In other words, the bond markets are large. How large are they? The answer to this question depends on one's definition of size.

Figure 1 shows that an astonishingly large number, approximately 1.5 million, corporate and municipal bonds are outstanding. The vast majority of these are municipal bonds, which are typically issued in serial offerings consisting of a set of up to 20 (or more) bonds issued at the same time with different maturities. Thus, the number of bonds dwarfs the number of equities.

In terms of total dollar amounts outstanding, Fig. 1 shows that US corporate and municipal bond markets combined are roughly half the size of the US equity markets. The average daily trading volume in these bond markets is about $36 billion, which is about 1/3 of the average daily trading volume of $115 billion in the equity markets. While the discussion of the microstructure of the Treasury bond markets is left to ► Treasury Market, Microstructure of the U.S., it is worth noting that total US bond market trading volume (corporate, municipal, and Treasury) exceeds US equity market trading volume. Thus, no matter what measure is used, it is apparent that the bond markets offer important sources of capital for issuers and securities for investors.

The remainder of this article proceeds as follows. Section "Introduction" provides a historical overview of the US corporate and municipal bond markets. Section "Early Corporate and Municipal Bond Market Microstructure Research" through "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" review the significant contributions to the bond



Sources: SIFMA, Board of Governors of the Federal Reserve System, NYSE, AMEX, Nasdaq, OTCBB, NASD

Corporate and Municipal Bond Market Microstructure in the U.S., Figure 1
Comparison US municipal corporate and bond markets with US equity markets

market microstructure literature. Section "Early Corporate and Municipal Bond Market Microstructure Research" reviews the early corporate and municipal bond market microstructure research. Section "Fixed Income Pricing Service (FIPS) Research" reviews the research enabled by the National Association of Securities Dealer's (NASD's) Fixed Income Pricing Service that began in May 1994. Section "Municipal Bond Market Research" reviews the municipal bond market research enabled by the Municipal Securities Rulemaking Board's (MSRB's) transaction data. Section "Transaction Reporting and Compliance Engine (TRACE) Research"reviews the research enabled by the NASD's Transaction Reporting and Compliance Engine (TRACE) system that began in July 2002. Section "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" provides examples of how bond market microstructure research is linked to other areas of finance and economics research.

Introduction

Today, virtually all US corporate and municipal bond trading occurs in over the-counter (OTC) dealer markets with transparent prices. But, that was not always the case. In the early 20th century there were active and transparent markets for both corporate bonds and municipal bonds on the New York Stock Exchange (NYSE). Then, bond trading began migrating to opaque OTC dealer markets. In the late 20th century, post-trade transparency was added to the both the corporate and municipal bond OTC markets.

What factors are responsible for the evolution of the bond markets over the past century? What caused the migration of trading in the early 20th century? How (and why) was post-trade transparency added to the bond markets in the late 20th century? The brief history of US corporate and municipal bond markets below provides answers to these questions.

The Early 20th Century

Biais and Green [9] provide a fascinating historical overview of the US corporate and municipal bond markets. Early 20th century NYSE bond trading took place among the "bond crowd". Bond trading originally took place in the same trading room as stock trading, with the bond crowd organizing around three trading booths in the "bond corner" of the Exchange. In 1928, the NYSE opened a separate trading room, the "bond room", in response to increases in trading volumes. Trading in the bond room was separated into four different crowds. US corporate and municipal bonds were traded in either the "active" crowd or the "inactive" crowd. The inactive crowd was also known as the "cabinet" crowd because bond orders were written on slips of paper and filed in the bond cabinets. Foreign bonds and Government securities each had their own bond crowds. A small number of active bonds were traded on the floor in an open outcry market.

NYSE bond trading was "order-driven". The exchange collected, posted, and matched public customer orders. Public investors paid commissions to brokers to facilitate their NYSE bond trades. All NYSE bond brokers could observe the book of available orders and the recent trades, and inform their customers about them. Thus, NYSE bond trading enjoyed a very high level of "pre-trade transparency" and "post-trade transparency". Pre-trade transparency refers to the dissemination of information about trading interests. Pre-trade information can include price (bid and ask) and depth quotations, as well as limit order prices and sizes. Post-trade transparency refers to the dissemination of information about past trades. While post-trade information includes not only prices, such as trade execution times and volumes, post-trade transparency in the bond markets is sometimes referred to as simply "price transparency". Madhavan [38] and Harris [24] provide excellent discussions all the different dimensions of transparency as well as the related market microstructure literature.

In the late 1920s, municipal bond trading migrated to the over the-counter (OTC) market. Corporate bond trading migrated to the OTC market in the 1940s. Biais and Green [9] examine a number of potential explanations for the decline in municipal and corporate bond trading on the NYSE. They find that the decline of exchange trading in bonds was not due to a decline in the supply of bonds outstanding or a decline in listings in response to costly rules and regulations promulgated by the newly created SEC.

Biais and Green [9] find that the migration of bond trading from the NYSE to the OTC markets coincided with changes in the investor base. In the late 1920s, retail investor interest in municipal bonds waned, as they became more attracted to the higher returns on equities. As retail interest in municipal bonds waned, institutions became the dominant investor in the market. During the 1940s, a similar shift in the relative importance of retail investors and institutional investors occurred in the corporate bond market. Biais and Green [9] conclude that the migration of bond trading from the NYSE to the OTC markets was an evolution in response to the changing investor base.

Biais and Green [9] provide evidence that institutions fared better in OTC bond markets and argue that the dealers were happy to accommodate this new class of dominant investors. Because liquidity was no longer concentrated on a centralized transparent exchange, retail investors were effectively forced into trading with dealers in these decentralized opaque OTC markets. Not surprisingly, retail investors faredjt worse in these markets. Both municipal and corporate bond transaction costs increased significantly for retail investors.

The Late 20th Century and Early 21st Century

While the most significant change in the bond markets in the early 20th century was a migration of trading from the exchange to OTC, the most significant change in the late 20th century was the introduction of price transparency. Unlike trading migration, bond market transparency was not caused by market forces. Rather, transparency was added to the bond markets by direct regulatory intervention.

The Municipal Securities Rulemaking Board (MSRB) introduced price transparency to the municipal bond market. The MSRB was created by Congress in 1975 as the self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their municipal bond transactions.

The MSRB began publicly disseminating municipal bond price information in January 1995. "Interdealer Daily Reports" provided statistics on total interdealer market activity reported for the previous day, as well as information about price and volume for each security that was "frequently traded" on that day. The MSRB defined frequently traded securities to be securities with four or more interdealer transactions on a particular day. The Interdealer Daily Report included the total par value traded, the daily high and low price, and the average price of trades having a par value between $100,000 and $1 million for each frequently traded issue. Transaction price information on securities with three or fewer interdealer transactions on a particular day ("infrequently traded" securities) was not disseminated.

In August 1998, the MSRB began producing "Combined Daily Reports". The Combined Daily Reports merged information from customer and interdealer transactions and provided daily high, low, and average prices for frequently traded securities of municipal securities on a one-day delayed basis. The frequently traded threshold was four transactions per day, taking into account both customer and interdealer transactions.

In January 2000, the MSRB began publicly disseminating transaction details on individual trades in frequently traded securities through "Daily Transaction Reports". Trade information on infrequently traded securities was still not disseminated until October 2000, when the MSRB began producing "Monthly Comprehensive Reports". These reports provided information on a one-month delayed basis for all transactions from the previous month, including infrequently traded issues.

By June 2003, the MSRB was publicly disseminating transaction details for all trades in all securities (frequently traded and infrequently traded) on a one-day lag basis through "T+1 Daily Reports". In January 2005, the MSRB began disseminating prices on a real-time basis through its Real-Time Transaction Reporting System (RTTRS or TRS).

The National Association of Securities Dealers (NASD) introduced price transparency to the corporate bond market. NASD (now FINRA) is the self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. Regulatory concern for price transparency spiked in the late 1980s and early 1990s.

At that time, the high-yield corporate bond market faced unprecedented instability, highlighted by insider trading scandals and the ultimate collapse of the dominant dealer and underwriter Drexel Burnham Lambert. Concern over future market instability, along with the recognition of a need for better monitoring, led to regulatory intervention that provided a small degree of transparency in this market segment. The Fixed Income Pricing Service (FIPS) began in 1994. FIPS was the result of the SEC encouraging NASD to develop an electronic reporting and dissemination facility for non-convertible high-yield corporate bonds.

But, FIPS only provided partial transparency for this particular segment of the corporate bond market. While every trade in FIPS-eligible bonds was reported to FIPS, only summary information on a small subset (50 bonds) of the most active bonds was disseminated to the public. Alexander, Edwards, and Ferri [2] point out that some members of the SEC staff at that time feared that adding price transparency to less active bonds could possibly harm the market. FIPS added both pre-trade transparency and post-trade transparency to the market by disseminating quotations and hourly trade summaries, respectively. The hourly trade summaries contained high and low prices as well as total trading volume.

Many bond market participants and some SEC staff felt that FIPS added a sufficient amount of transparency to the corporate bond market. SEC Chairman Arthur Levitt disagreed. In 1998, he gave a speech entitled *The Importance of Transparency in America's Debt Market* in which he famously quipped "The sad truth is that investors in the corporate bond market do not enjoy the same access to information as a car buyer or a homebuyer or, dare I say, a fruit buyer". To address the lack of price transparency in the corporate bond market, he called on NASD to take several related actions. He called on NASD to adopt rules requiring dealers to report all corporate bond transactions; to develop a system to receive all corporate bond transaction information; to create a database of the transactions, and in conjunction, create a surveillance program to better detect fraud in corporate bonds; and, to disseminate the bond transaction prices to the public in order to help them make better investment decisions.

NASD responded by developing the Transaction Reporting and Compliance Engine (TRACE) system, which began operation in July 2002. Corporate bond dealers were required to report all transaction in TRACE-eligible securities. TRACE-eligible securities included investment grade and high-yield debt, convertible and non-convertible debt, and publicly issued debt and privately issued (Rule 144A) debt. While all TRACE-eligible transactions were reported to TRACE from the beginning of its operation, the dissemination of the trade information was phased-in over time. The phase-in approach was adopted by NASD, and approved by the SEC, because of industry concerns that adding transparency to the bond market would somehow harm liquidity.

The TRACE phase-in approach began with the dissemination of trade information on the largest, highest-rated bonds first. Price transparency was introduced to smaller and lower-rated bonds over time. By February 2005, prices were transparent on effectively 99% of trades, and by the end of that year, pricing information on all TRACE-eligible trades was being disseminated on a real-time basis to the public.

By the beginning of the 21st century, investors (and market microstructure researchers) were able to access an unprecedented amount of information about the OTC municipal and corporate bond markets from the post-trade transparency brought by TRS and TRACE, respectively. It is worth noting that bond trading never completely migrated to the OTC markets. The NYSE continues to list and trade some bonds. The NYSE developed the Automated Bond System (ABS), a computerized limit-order market for bonds, in an effort to encourage the migration of trading back to the exchange. The displayed public limit orders on ABS provided pre-trade transparency for some bonds.

However, the vast majority of bonds are not listed on the NYSE ABS or any other exchange, so all of the trading in these bonds occurs in the OTC markets. Moreover, for many of the bonds that are listed on the NYSE ABS, a majority of their trades still occur over the-counter. Therefore, the early 21st century bond markets can be characterized as dealer markets with a high degree of post-trade transparency, but with virtually no pre-trade transparency. It remains to be seen whether market forces, regulatory initiatives, or some combination of the two will eventually lead to the introduction of some form of pre-trade transparency, the emergence of bond market-makers, and/or a migration of trading back to an order-driven market in the US corporate and municipal bond markets.

Early Corporate and Municipal Bond Market Microstructure Research

Early bond market microstructure researchers were forced to rely on data sources that covered only certain segments of the market because comprehensive bond market data simply did not exist. Bond dealers kept their own trading records because there was no central reporting facility. Bond dealers were not required to publicly disseminate their trades. Researchers, as well as investors and regulators, were able to see snapshots of parts of the bond markets, but no one was able to see the complete picture.

But, even with the data limitations, early bond market researchers found creative ways to tease out useful information. Their initial findings shed the first light on the opaque bond markets. For example, Schultz [42] provides indirect evidence that higher bond market trading costs may be attributable to the lack of transparency. Variants of their original empirical methods continue to be used by more recent researchers.

Any encyclopedic article on corporate bond market microstructure research would be incomplete if it did not mention the efforts of the Fixed Income Research Program (FIRP), and more importantly its founder Professor Arthur Warga, in promoting early bond market research. Art Warga's influence on the development of the bond market microstructure literature extends beyond his own research. He collected, consolidated, cleaned, and organized various fragmented sources of bond market data to create the Fixed Income Securities Database (FISD), which he made accessible to academic and regulatory researchers. Many within the market microstructure research community informally refer to the FISD as simply the "Warga database".

Warga (1991)

Warga [44] uses an econometric model to investigate bond pricing discrepancies that arise when researchers (and commercial bond pricing services) use data from the two different sources that were generally available in 1990. The one source was exchange data in the form of actual transaction prices from the NYSE Automated Bond System (ABS). The other source was OTC dealer data in the form trader-quoted prices.

Warga [44] denotes the unobserved, true value of bond i as P_i^* and the unobserved, true bid-ask spread as BA_i. He assumes that P_i^* is the midpoint of BA_i. He also assumes that the prices/quotes observed in both markets are unbiased in the sense that they deviate from the true unobserved prices/quotes by a random error term. Then, for month-end NYSE transaction prices (P_{NY}):

$$P_{NY_i} = P_i^* + u_i \ ,$$

and, for month-end Lehman Brothers bid quotes (P_B):

$$P_{B_i} = P_i^* - \frac{1}{2} BA_i + \zeta_i \ .$$

Combining these two equations and letting $\varepsilon_\iota = \zeta_\iota + \mu_\iota$ yields:

$$P_{B_i} - P_{NY_i} = -\frac{1}{2} BA_i + \varepsilon_i \ .$$

Squaring both sides results in:

$$\left(P_{B_i} - P_{NY_i}\right)^2 = \frac{1}{4} (BA_i)^2 - (BA_i)\,\varepsilon_i + \varepsilon_i^2 \ .$$

Assuming the random error terms are orthogonal to prices/quotes, the expected squared price discrepancies is:

$$E\left[\left(P_{B_i} - P_{NY_i}\right)^2\right] = \frac{1}{4} (BA_i)^2 + \sigma_{\varepsilon i}^2 \ ,$$

where σ_ε^2 equals the variance of the discrepancy.

Warga [44] regresses the squared price discrepancies on six observable liquidity-related variables – bond rating, duration, NYSE dollar trading volume, bond age, issue amount outstanding, and the time of trade of the last trade price on the NYSE – with the following equation:

$$\begin{aligned}\left(P_{B_i} - P_{NY_i}\right)^2 &= \alpha_0 + \alpha_1 MOODYS + \alpha_2 DURTN \\ &\quad + \alpha_3 OUTSTD + \alpha_4 DVOL + \alpha_5 AGE + \alpha_6 TIME + \omega \ .\end{aligned}$$

Warga [44] finds that squared discrepancies are larger for bonds with lower credit ratings, higher duration, smaller issue sizes, lower trading volume, and trade prices that occurred earlier in the day. While he finds that these variables are capable of explaining some of the observed variation in price discrepancies, he also concludes that commingling exchange and dealer bond pricing data does not induce any biases.

Hong and Warga (2000)

Hong and Warga [29] use a benchmark method to estimate average daily effective spreads with newly available trade data. They obtain exchange market data from the NYSE ABS and OTC dealer market data from the Capital

Access International (CAI) database. CAI obtains trading data on insurance companies, mutual funds, and pension funds from various regulatory filings.

They calculate the effective spread for a given bond on a given day as the dollar-volume-weighted average price transacted at the ask minus the dollar-volume-weighted average price transacted at the bid:

$$\sum_{i=1}^{N} P_i^{\mathrm{A}} W_i^{\mathrm{A}} - \sum_{j=1}^{M} P_j^{\mathrm{B}} W_j^{\mathrm{B}} ,$$

where P_i^{A} is the price of transaction i occurring at the ask, W_i^{A} is the dollar-value weight of transaction i, and N is the number of transactions occurring at the ask for a given bond on a given day. Similarly, P_j^{B} is the price of transaction j occurring at the bid, W_j^{B} is the dollar-value weight of transaction j, and M is the number of transactions occurring at the bid for a given bond on a given day.

Hong and Warga [29] find that the average daily effective spreads for corporate bond transactions occurring on the NYSE ABS that involve at least 10 bonds is about $0.21 for investment grade bonds and about $0.19 for high-yield bonds. For corporate bond trades occurring in the OTC dealer market, they find that the average daily effective spreads is about $0.13 for investment grade bonds and about $0.19 for high-yield bonds. They note that these spread estimates are smaller than previous estimates based on data from an earlier period, which is consistent with evidence that corporate bond spreads may have declined over time.

Hong and Warga [29] also find that OTC dealer market spreads exhibit much larger dispersion than NYSE ABS spreads. The standard deviations of daily effective spreads for the dealer market are two to three times larger than those for the exchange market. This result suggests that investors, particularly uninformed retail investors, could benefit from more transparency in the OTC markets.

Schultz (2001)

Schultz [42] also uses CAI institutional trade data. He develops an econometric model similar to Warga [44] to estimate average round trip corporate bond trading costs from institutional trade data and estimated contemporaneous bid quotes.

Schultz [42] estimates daily corporate bond bid quotes from the month-end bid quotes available from the Warga database. He develops a threestep estimation procedure that uses daily Treasury bond bid quotes, based on the observation that most of the day-to-day changes in investment grade corporate bond prices are explained by changes in the level of risk-free interest rates.

The first step is to calculate a predicted month-end quote for each corporate bond by taking its previous month-end quote and multiplying it by the change in the price of Treasury bonds of similar duration. The second step is to subtract the predicted month-end quote from the actual month-end quote. This calculation yields the monthly pricing error from predicting that the change in the corporate bond prices is exactly the same as the change in Treasury bond prices. The monthly pricing error is converted to an average daily pricing error by dividing it by the number of trading days in the month. The third step is to estimate the bid quote for a particular within-month trade date by starting with the previous end-of-month quote and adding on the average daily pricing error times the number trading days since the previous month end.

Schultz [42] finds that his bid quote estimates are accurate for investment grade bonds, but not for high-yield bonds. This is not surprising, since changes in high-yield prices are more often due to changes in firm-specific factor than changes in Treasury bond prices. Therefore, he does not attempt to estimate trading costs for high-yield bonds with this methodology.

For investment grade bonds, Schultz [42] estimates round-trip transactions costs by regressing the difference between the CAI trade prices and his estimate of the contemporaneous bid quote on a dummy variable that takes the value of 1 for buys and 0 for sells:

$$\Delta_i = \alpha_0 + \alpha_1 D_i^{\mathrm{Buy}} + \varepsilon_i ,$$

where Δ_i is the price of trade i minus estimated bid price and D_i^{Buy} equals one if trade i is a buy and zero otherwise. The coefficient α_i is an estimate of the average round-trip transaction costs. His estimate of the average round-trip transaction costs across all trades is about $0.27 per $100 of par value.

Schultz [42] also examines the determinants of corporate bond trading costs with the following regression:

$$\begin{aligned}\Delta_i = \alpha_0 &+ \alpha_1 D_i^{\mathrm{Buy}} + \alpha_2 S_i + \alpha_3 D_i^{\mathrm{Inst}} + \alpha_4 D_i^{\mathrm{Deal}} \\ &+ \alpha_5 D_i^{\mathrm{Inst}} D_i^{\mathrm{Deal}} + \alpha_6 D_i^{\mathrm{Inst}} S_i + \alpha_7 D_i^{\mathrm{Deal}} S_i + \varepsilon_i ,\end{aligned}$$

where S_i is the signed (positive for buys, negative for sells) natural logarithm of the dollar trade size, D_i^{Inst} is a dummy variable that takes a value of one for buys and negative one for sells by one of the 20 most active institutions and a value of zero otherwise, and D_i^{Deal} is a dummy variable that takes a value of one for buys and negative one for sells if the trade involves one of the 12 active dealers and a value of zero otherwise. For the interactive term for active institutions and dealers, $D_i^{\text{Inst}} D_i^{\text{Deal}}$, the product of the dummies is positive for buys and negative for sells when the trade involves both an active institution and an active dealer and a value of zero otherwise.

Schultz [42] finds that institutional corporate bond trading costs decline with trade size. He does not find any evidence that trading costs are related to credit rating. But, this is not surprising given the fact that his analysis is limited to the four investment grade rating classes (Aaa, Aa, A, Baa).

Schultz [42] finds that trading costs are lower when a large bond dealer is used. In other words, small bond dealers charge more than large ones. He also finds that inactive institutions pay more than twice as much as active institutions to trade the same bonds. Schultz [42] attributes this result to the lack of transparency in the corporate bond market during his sample period. In an opaque market, obtaining price information is costly, so only active institutions will find it worthwhile to bear them.

Chakravarty and Sarkar (2003)

Chakravarty and Sarkar [12] use CAI data and benchmark methods to calculate "traded bid ask-spreads" over one-day, two-day, and five-day windows in the corporate, municipal, and Treasury bond markets. Similar to Hong and Warga [29], they define the traded bid-ask spread per day as the difference between its mean daily selling price and its mean daily buying price.

To check the sensitivity of their estimates to the requirement of one buy trade and one sell trade for each bond day, Chakravarty and Sarkar [12] calculate spreads over non-overlapping two-day and five-day windows. Their two-day traded bid-ask spread is calculated as the difference between the two-day means of the selling prices and the buying prices.

Chakravarty and Sarkar [12] find that the mean traded bid-ask spread per day per $100 par value is $0.21 for corporate bonds, $0.23 for municipal bonds, and $0.08 for Treasury bonds. In all three markets, they find that spreads increase with longer time-to-maturity, lower credit ratings, and lower trading volume. These results suggest that spreads are positively related to interest rate risk and credit risk, and negatively related to trading activity. For corporate bonds, Chakravarty and Sarkar [12] find that spreads increase with age.

Chakravarty and Sarkar [12] pool observations across all three bond markets for cross-market comparisons. After controlling for credit risk, Chakravarty and Sarkar [12] find no significant difference in the spreads of corporate bonds and Treasury bonds, but they find that municipal bonds have higher spreads.

Fixed Income Pricing Service (FIPS) Research

FIPS provided new price and volume data that allowed market microstructure researchers to conduct studies that were not previously possible. Alexander, Edwards, and Ferri [1] use FIPS volume data to test various hypotheses about bond liquidity. Alexander, Edwards, and Ferri [2] use FIPS returns and equity returns to tease out new evidence on agency conflicts between stockholders and bondholders. Hotchkiss and Ronen [31] use FIPS returns and equity returns to examine the relative informational efficiency of the corporate bond market. Somewhat surprisingly, they find that that informational efficiency of the corporate bond market is similar to the stock market.

Alexander, Edwards, and Ferri (2000a)

Alexander, Edwards, and Ferri [1] examine the determinants of trading volume of FIPS high-yield bonds using a pooled time-series cross-sectional approach. They use the following linear specification:

$$\begin{aligned} \mathit{Trading\ Volume}_{it} &= \beta_0 + \beta_1 \mathrm{Ln}\,(\mathit{Size}_{it}) + \beta_2 \mathit{Age}_{it} \\ &\quad + \beta_3 \mathit{Private\ Equity}_{it} + \beta_4 \mathit{Credit\ Rating}_{it} \\ &\quad + \beta_5 \mathit{Duration}_{it} + \beta_6 \mathit{Price\ Variability} + \varepsilon_{it}\,, \end{aligned}$$

where the dependent variable, *Trading Volume*, is measured in three different ways for each bond i in each month t. The three trading volume measures are the natural log of the average daily number of trades (Ln*(Trades)*), the natural log of the average daily number of bonds traded (Ln*(Bonds)*), and average daily turnover (*Turnover*). Ln*(Size)* is the natural logarithm of the issue's par value outstanding, *Age* is a dummy variable equal to one if the issue has been outstanding for less than two years, *Private Equity* is a dummy variable equal to one if the issue has no public equity outstanding in any part of the month, *Credit Rating* is a dummy variable equal to one if the issue is rated below B- by Standard & Poor's at any point during the month, *Duration* is the bond's modified duration, and *Price Variability* is the monthly average absolute value of the daily percentage change in volume-weighted price.

They find consistent results for all three measures of trading volume. Larger issues and younger issues are more heavily traded. They point out that the age result extends earlier empirical results that found that the liquidity of Treasury securities drops off a few months after issuance.

Alexander, Edwards, and Ferri [1] also find that bonds of firms without public equity trade more frequently than bonds of firms with public equity. This last finding is inconsistent with a disclosure hypothesis that predicts that more relaxed disclosure rules for firms without public equity will lead lower liquidity, as measured by trading volume. However, it is consistent with the competing substitution hypothesis that predicts that high-yield bonds of private firms will attract trading volume that otherwise would have occurred in the equity.

Alexander, Edwards, and Ferri (2000b)

Alexander, Edwards, and Ferri [2] use equity data and FIPS bond data to investigate the relationship between a firm's stock return and its bond return. They examine the long-term co-movement between a firm's bond returns and its stock returns. They also examine stock and bond returns around events typically associated with agency conflicts to see whether their co-movements provide evidence of agency conflicts.

To examine the long-term co-movement between a firm's bond returns and its stock returns, Alexander, Edwards, and Ferri [2] use three regression approaches. The first approach is a time-series regression model:

$$\begin{aligned} RB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \beta_3 RBIND_{it} \\ + \beta_4 RBIND_{it-1} + \varepsilon_{it} \ , \end{aligned}$$

where RB_{it} is the bond return for firm i on day t, XRS is the current (t) and lagged ($t-1$) excess stock return, and $RBIND$ is the current (t) and lagged ($t-1$) high-yield bond index return, and ε_{it} is the residual bond return for firm i on day t. The second approach is a pooled time-series cross-sectional model that uses the regression equation above and follows the pooling technique of Greene (1993). The third approach is a cross-sectional regression model that follows the approach of Fama and MacBeth. For each sample day, the excess bond returns are regressed on the current and lagged excess stock returns:

$$XRB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \varepsilon_{it} \ .$$

The estimates of β_1 in each of the three regressions show whether the stock and bond returns tend to co-move together (positive), in the opposite direction (negative), or not at all (insignificant). Alexander, Edwards, and Ferri [2] find that all three regressions produce similar results. The β_1 estimates are positive and statistically significant, indicating that excess bond returns are positively correlated with excess stock returns. But, Alexander, Edwards, and Ferri [2] point out the that the magnitudes of the coefficients suggest that the correlation is economically small.

To examine the behavior of stock and bond returns around events typically associated with agency conflicts, Alexander, Edwards, and Ferri [2] look at cumulative excess stock and bond returns around announcements of corporate events that are typically associated with wealth transfers from bondholders to stockholders, or vice versa. Events include debt issuances and redemptions, stock issuances and repurchases, dividend changes, new credit agreements, and others. They use Wilcoxon rank-sum tests to determine whether the means of the cumulative excess bond returns around potentially wealth-transferring events are significantly different from the returns otherwise. They find that the means are significantly different and that the mean cumulative excess bond returns around the wealth-transferring events is negative, while the returns are at other times are positive.

Thus, Alexander, Edwards, and Ferri [2] show that wealth-transferring corporate events (from bondholders to stockholders, or vice versa) can cause a firm's bond returns to diverge from its typical positive (weak) co-movement with its stock returns. In addition, they point out that this result is a likely factor in the weak long-term time-series correlations observed between stock and bond returns.

Hotchkiss and Ronen (2002)

Hotchkiss and Ronen [31] use FIPS data for 55 high-yield bonds to examine the informational efficiency of the corporate bond market relative to the market for the underlying stock. They find that stocks do not lead bonds in reflecting firm-specific information. They also find that pricing errors for bonds are no worse than for the underlying stocks, even on an intraday level.

Hotchkiss and Ronen [31] use a vector autoregression (VAR) approach applied to daily and hourly returns with a return-generating process that includes an interest rate risk factor and an equity market (systematic) risk factor:

$$\begin{aligned} RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{L}} RL_{t-i} \\ + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t \ , \end{aligned}$$

where RB_t is the FIPS bond portfolio return, RL_t is the Lehman Intermediate Government Bond Index return, and RM_t is the S&P 500 Index return. The number of lags for the bond, interest rate, and stock returns are $nb = 3$, $ni = 0$, and $ns = 4$, respectively. Hotchkiss and Ronen [31] include lagged bond returns (RB_{t-i}) to consider autocorrelation-adjusted bond returns. They also consider a specification that replaces the Lehman index return with the default risk-free return, RD_t, as interest rate factor:

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t \,.$$

Finally, they add the underlying stock (firm-specific) return, RS_t :

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{S}} RS_{t-i} + \varepsilon_t \,.$$

With these three regressions, Hotchkiss and Ronen [31] find that high-yield bond returns exhibit a very strong interest rate risk component and that this component is significantly greater for higher-rated bonds. They also find that high-yield bond returns exhibit a very strong systematic risk component and that this component is slightly weaker for higher-rated bonds.

To test whether stock returns lead bond returns, Hotchkiss and Ronen [31] conduct Granger causality tests at the daily and hourly levels. They estimate the VAR for the variable set $z_t = [RB_t, RS_t]'$ using the specification:

$$z_t = B_1 z_{t-j} + B_2 z_{t-j} + \mu_t \,,$$

where RB_t is the bond return and RS_t is the stock return, for day (hour) t, B_i are conformable matrices, and μ_t is a disturbance vector. To test whether stock returns Granger cause bond returns they estimate the following bivariate VAR model using ordinary least squares (OLS):

$$RB_t = c_1 + \sum_{i=1}^{j} a_i RB_{t-i} + \sum_{i=1}^{j} b_i RS_{t-i} + \nu_{1,t} \,,$$

where c is a constant, as and bs are coefficients, ν_t is the disturbance vector, and j is the lag length. The null hypothesis is that stock returns do not Granger cause bond returns, or that $H_0 = [b_i] = 0$, for all i. Tests of whether bond returns Granger cause stock returns are conducted in a similar way. F-tests indicate that lagged stock returns are not significant in explaining bond returns. Thus, stocks do not lead bonds in reflecting firm-specific information. The Granger causality test results also indicate that lagged bond returns are not significant in explaining stock returns.

Hotchkiss and Ronen's [31] interpretation of the Granger causality test results is that the contemporaneous correlations between stock returns and bond returns are best described as a joint reactions to common factors. This motivates an additional investigation of the comparative reaction of stocks and bonds to firm-specific information. To conduct this investigation, they examine how quickly firm-specific information contained in earnings announcements are incorporated into bond prices relative to stock prices. First, they compare reported earnings to the median of analysts' earnings forecasts and calculate the log forecast error:

$$FE_i = \ln \left(A_i / F_i \right) \,,$$

where FE_i is the log forecast error for firm i, A_i is the announced earnings per share, and F_i is the forecast earnings per share. Next, they run the following regressions to examine whether earnings information is reflected in bond returns or stock returns:

$$\begin{aligned} RB_{[-1,t]} &= \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon \\ RS_{[-1,t]} &= \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon \,, \end{aligned}$$

where RB and RS are the bond and stock returns, respectively, for the period starting at day (hour) -1 prior to the announcement and ending at day $+7$ (hour $+14$) after the announcement, and RM is the market (S&P 500 Index) return. Both the daily and hourly regression results indicate that all information is quickly impounded into both bond prices and stock prices.

Finally, Hotchkiss and Ronen [31] compare the market quality for the high-yield FIPS bonds to the underlying stocks by examining whether price errors of different magnitudes are associated with the different markets. The estimate the following market quality measure:

$$MQ_i = 1 - 2 * \left(\sigma_{si}^2 / \sigma_{Ri}^2 \right) \,,$$

where σ_{si}^2 is the variance of the pricing error described in Hasbrouck [27] and σ_{Ri}^2 is the variance of the return. The intuitive interpretation of this measure is the proportion of the total return variance that is due to fundamental variance. In general, they find that the market quality measure for bonds is no worse than for the underlying stocks.

Municipal Bond Market Research

With the MSRB's introduction of central reporting and price transparency to the municipal bond market, microstructure researchers were able to make use of a comprehensive source quality transaction-level municipal bond data for the first time. This new data provided researchers the opportunity to develop new methods, examine existing microstructure issues in greater detail, and identify new avenues of research.

Two prominent municipal bond studies, Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22], develop and use very different methods to examine various economic aspects of trading in the municipal bond market. These two studies provide independent sets of similar and robust results that support two important conclusions related to retail investors. The first is that municipal bonds are expensive for retail investors to trade. Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22] both find that, unlike in equity markets, municipal bond trading costs decrease with trade size.

The second is that "complexity" is costly for retail investors. Harris and Piwowar [26] find that "instrument complexity" makes municipal bonds more expensive to trade. Instrument complexity is measured in terms of attached features, such as calls, puts, sinking funds, credit enhancement, nonstandard interest payment frequencies, and nonstandard interest accrual methods. Green, Hollifield, and Schurhoff [22] find that "deal complexity" also increases trading costs. Bond dealers charge higher markups on more difficult trades.

Harris and Piwowar (2006)

Harris and Piwowar [26] estimate municipal bond trading costs using an econometric model. They denote the unobserved "true value" of the bond at the time t asV_t and assume that the price of a trade, P_t, is equal to V_t plus or minus a price concession that depends on whether the trade is buyer-initiated or seller-initiated. The absolute customer transaction cost, $c(S_t)$, is estimated as the effective half-spread, measured as a percentage of the price.

I_t^{D} is an indicator variable that takes a value of 1 if the trade was an interdealer trade or 0 if the trade was a customer trade. Q_t is an indicator variable that takes a value of 1 if the customer was a buyer, -1 if the customer was a seller, or 0 if it was an interdealer trade. This results in:

$$\begin{aligned} P_t &= V_t + Q_t P_t c(S_t) + I_t^{\mathrm{D}} P_t \delta_t \\ &= V_t \left(1 + \frac{Q_t P_t c(S_t) + I_t^{\mathrm{D}} P_t \delta_t}{V_t}\right) . \end{aligned}$$

The continuously compounded bond price and "true value" returns between trades t and s, r_{ts}^{P} and r_{ts}^{V} respectively, are found by taking logs of both sides, making two small approximations, and subtracting the same expression for trade s:

$$r_{ts}^{\mathrm{P}} = r_{ts}^{\mathrm{V}} + Q_t c(S_t) - Q_s c\,(S_s) + I_t^{\mathrm{D}} \delta_t - I_s^{\mathrm{D}} \delta_s \,.$$

The "true value" return r_{ts}^{V} is represented with a factor model by decomposing it into the linear sum of a time drift, a short-term municipal bond factor return, a long-term municipal bond factor return, and a bond-specific valuation factor, ε_{ts}:

$$\begin{aligned} r_{ts}^{\mathrm{V}} &= Days_{ts} \left(5\% - CouponRate\right) \\ &\qquad + \beta_{\mathrm{Avg}} SLAvg_{ts}^{+} \beta_{\mathrm{Dif}} SLDif_{ts}^{+} \varepsilon_{ts} \,, \end{aligned}$$

where $Days_{ts}$ counts the number of calendar days between trades t and s, $CouponRate$ is the bond coupon rate. $SLAvg_{ts}$ and $SLDif_{ts}$ are the average and difference, respectively, of continuously compounded short- and long-duration factor returns between trades t and s. The first term models the continuously compounded bond price return that traders expect when interest rates are constant and the bond's coupon interest rate differs from a notional five percent bond, the median coupon rate in their sample. The two index returns model municipal bond value changes due to changes in interest rates and tax-exempt yield spreads. Harris and Piwowar [26] use repeat sales methods to estimate these indices. They assume that the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma_{\varepsilon_{ts}}^2 = N_{ts}^{\mathrm{Sessions}} \sigma_{\mathrm{Sessions}}^2$$

where $N_{ts}^{\mathrm{Sessions}}$ is the total number of full and partial trading sessions between trades t and s.

To model customer transaction costs, Harris and Piwowar [26] consider several alternative functional forms that are flexible enough to model very high average trading costs for small trade sizes and very low average trading costs for large trade sizes. Harris and Piwowar [26] choose following parsimonious expression:

$$c(S_t) = c_0 + c_1 \frac{1}{S_t} + c_2 \log S_t + \kappa_t \,,$$

where the first three terms specify the cost function that represents average trade costs and κ_t represents the unexplained variation in the observed customer trading costs. The constant term allows total transaction costs to grow in proportion to size. The second term captures fixed costs per trade and the third term allows the costs per bond to vary by size.

The Harris and Piwowar [26] time-series estimation model is obtained by combining the last four equations:

$$\begin{aligned} r^{\mathrm{P}}_{ts} - Days_{ts}\left(5\% - CouponRate\right) &= c_0\left(Q_t - Q_s\right) \\ &+ c_1\left(Q_t\frac{1}{S_t} - Q_s\frac{1}{S_s}\right) + c_2\left(Q_t\log S_t - Q_s\log S_s\right) \\ &+ \beta_{\mathrm{SLAvg}} SLAvg^{+}_{ts}\beta_{\mathrm{SLDif}} SLDif^{+}_{ts}\,\eta_{ts}\ , \end{aligned}$$

where the expression for the regression term, η_{ts}, is given by:

$$\eta_{ts} = \varepsilon_{ts} + Q_t\kappa_t - Q_s\kappa_s + I^{\mathrm{D}}_t\delta_t - I^{\mathrm{D}}_s\delta_s\ .$$

The mean of the error term is zero and its variance is:

$$\sigma^2_{ts} = N^{\mathrm{Sessions}}_{ts}\sigma^2_{\mathrm{Sessions}} + D_{ts}\sigma^2_{\delta} + \left(2 - D_{ts}\right)\sigma^2_{\kappa}\ ,$$

where D_{ts} represents the number (0, 1, or 2) of interdealer trades involved in trades t and s. For each bond, Harris and Piwowar [26] separately estimate their time-series transaction cost estimation model using an iterated least squares method, with the weight given by the inverse of the estimates of σ^2_{ts}. For a wide range of trade sizes, they calculate weighted cross-sectional mean cost estimates across all municipal bonds. Each bond's weight is given by the inverse of its estimation error variance at that trade size.

Harris and Piwowar [26] find that retail-size municipal bond trades are substantially more expensive than similar-sized equity trades. Average effective spreads in municipal bonds are almost 2% for representative retail-size trades ($20,000). They point out that this is the equivalent of almost 4 months of total annual return for a bond with a 6% yield-to-maturity.

Harris and Piwowar [26] also find that retail-size municipal bond trades are more expensive than institutional-size trades. Unlike in equities, municipal bond transaction costs decrease with trade size. Harris and Piwowar [26] also find that, unlike in equities, municipal bond transaction costs do not depend on trade frequency. They attribute these results to the lack of price transparency in the municipal bond market during their sample period.

To investigate how estimated transaction costs vary across municipal bonds, Harris and Piwowar [26] conduct cross-sectional weighted least squares regressions for various trade sizes. The dependent variable is the estimated average transaction costs in a given municipal bond at a given trade size. The weight for each bond observation is given by the inverse of the estimation error variance of its cost estimate. The independent variables include measures of credit quality, age, and instrument complexity.

Harris and Piwowar [26] show that bond trading costs increase with credit risk, time to maturity, and time since issuance. They also find that trading costs increase with instrument complexity, and that retail investors are more adversely affected by instrument complexity than institutional investors. They conjecture that investors and issuers might benefit if simpler bonds were issued.

Green, Hollifield, and Schurhoff (2007a)

Green, Hollifield, and Schurhoff [22] focus on trades that can reasonably be assumed to represent two sides of a single intermediated transaction, and employ a structural model to decompose the cost faced by a customer into a portion that represents the cost the dealer incurs and a portion attributable to the dealer's market power. They formulate and estimate a simple structural bargaining model that allows them to estimate measures of dealer bargaining power and relate it to characteristics of the trades.

Green, Hollifield, and Schurhoff [22] use a theoretical model to seek evidence that the high costs of trading are due to dealer market power and to find out how the exercise of market power depends on the characteristics of the trade. They develop a simple theoretical model of the interaction between dealers and their customers in which the expected profits to the dealer reflect both the dealer's costs and his bargaining power relative to the customer. Both of these, in turn, can be parametrized as functions of observable variables, and estimated as a Stochastic Frontier Model. The dealer's cost is the stochastic frontier, which represents the expected mark-up the customer would obtain if dealers were always driven by their reservation values, as they would be if the provision of dealer services were perfectly competitive. The observed markup, expressed in excess returns over a municipal bond index, can be written as:

$$\begin{aligned} &\frac{p_i - p^*_i}{p^*_i} - R_{\mathrm{index},i} \\ &\qquad = \left[\frac{c\left(X_i,\theta\right)}{p^*_i} - E\left(R_{\mathrm{index},i}\middle| X_i\right)\right] + \varepsilon_i + \xi_i\ , \end{aligned}$$

where p_i is the dealer's selling price, p^*_i is the dealer's purchase price, and $R_{\mathrm{index},i}$ is the municipal bond market index return.

The first term on the right-hand side of the equation represents the dealer's costs in excess of the expected municipal bond index return, where X_i is a set of conditioning variables observable to the buyer and seller and θ is a set of parameters to be estimated. They refer to this term as the cost of intermediation.

The second and third terms capture how the observed markup can differ from the dealer's cost of intermediation.

The second term, ε_i, is a symmetric, normally-distributed error term:

$$\varepsilon_i \equiv \frac{e_i}{p_i^*} - \eta_i \,,$$

reflecting a zero-mean forecast error:

$$\eta_i = R_{\text{index},i} - E\left(R_{\text{index},i} \middle| X_i\right) \,.$$

The third term, ξ_i, is a one-sided, exponentially-distributed error term:

$$\xi \equiv \frac{\rho_i\left[E\left(p_i \middle| X_i\right) - c\left(X_i, \theta\right) - v_i\right]}{p_i^*} \,,$$

reflecting the distribution of sellers' reservation values (v_i) and dealer bargaining power.

Green, Hollifield, and Schurhoff [22] estimate restricted and unrestricted versions of the following regression model via maximum likelihood:

$$\frac{p_i - p_i^*}{p_i^*} - R_{\text{index},i} = \theta_0 + \sum_{l=1}^{L} \theta_l X_{il} + \varepsilon_i + \xi_i \,,$$

with $l = 1, \ldots, L$ conditioning variables. The residual ε_i is normally distributed with standard deviation $b_0 \Pi_{k=1}^{K} \mathrm{e}^{b_{ik} Z_k}$, with Z_{ik} for $k = 1, \ldots, K$ conditioning variables. The residual ξ_i is exponentially distributed with mean and standard deviation $a_0 \Pi_{k=1}^{K} \mathrm{e}^{a_{ik} Z_k}$. In the "market power" version of their model, all of the parameters are unrestricted. In the restricted ("no market power") model, all of the parameters on the one-sided error are constrained to zero: $a_0 = a_1 = \ldots = a_k = 0$.

The data used by Green, Hollifield, and Schurhoff [22] includes both customer trades and interdealer trades. But, because their data does not identify the specific broker-dealer associated with a given trade, they must infer their trades and profits indirectly by studying pairs of trades that appear to be direct exchanges of bonds through a single dealer. They assume that a customer buy transaction of a given size of a given bond that occurs within a very short time of customer sell transaction of the same size in the same bond are most likely related. The reasonableness of this assumption is confirmed by Harris and Piwowar [26], whose data contains dealer identities.

Green, Hollifield, and Schurhoff [22] find that municipal bond dealers earn lower average markups on larger trades, even though larger trades lead the dealers to bear more risk of losses. Their results suggest that municipal bond dealers exercise substantial market power, particularly in retail-sized transactions. Their measures of market power decrease in trade size and increase in variables that indicate the complexity of the trade for the dealer.

Transaction Reporting and Compliance Engine (TRACE) Research

TRACE not only brought unprecedented transparency to corporate bond market investors, it also provided an unprecedented opportunity for market microstructure researchers to examine new issues. Chief among them was the "natural experiment" of adding price transparency to an opaque market. Three prominent studies (collectively, "the TRACE studies") that examined the introduction of price transparency to the corporate bond market were Edwards et al. [17], Bessembinder, Maxwell, and Venkataraman [7], and Goldstein, Hotchkiss, and Sirri [20].

These TRACE studies were very complementary in terms of their contributions to the market microstructure literature. To understand the full impact of this collective research, it is important to remember that they were written at a time when many market participants and some regulators were concerned that public dissemination of bond pricing data might have an adverse impact on liquidity. Using different experimental designs and empirical methods, the TRACE studies produced similar results, conclusions, and implications for regulatory policymakers. Overall, the results in all three TRACE studies show that public investors benefit significantly from the introduction of price transparency.

Edwards et al. [17] estimate transaction costs for all corporate bonds that trade at least nine times between January 2003 and January 2005. Their TRACE data set includes all reported OTC trades in corporate bonds, whether transparent or not. Consistent with the results of Harris and Piwowar [26] for the municipal bond market, Edwards et al. [17] find that corporate bonds are expensive for retail investors to trade and that corporate bond transaction costs decrease significantly with trade size. They find that effective spreads in corporate bonds average 1.24% of the price of representative retail-sized trades ($20,000). They point out that this is the equivalent of over 2 months of the total annual return for a bond with a 6% yield to maturity, or 52 cents per share for a $40 stock. In cross-sectional tests, Edwards et al. [17] find that transaction costs are lower for highly rated bonds, recently issued bonds, and bonds close to maturity.

Edwards et al. [17] find that costs are lower for bonds with transparent trade prices, and they drop when the TRACE system starts to publicly disseminate their prices. Their results suggest that introduction of price transparency results in a drop in customer trading costs of at least 5 basis points (bps). In 2003, public investors traded approximately $2 trillion in bonds for which prices were not disseminated. If the prices for these bonds had been

TRACE-transparent, a quick back-of the-envelope calculation shows investors could have saved a minimum of $1 billion that year. Edwards et al. [17] point out that the $1 billion figure represents a lower bound for two reasons. First, because many unsophisticated investors were unaware that prices became available, and because learning how to use the price data takes, time, the long-run benefits are undoubtedly much greater. Second, they do not capture the initial reduction in trading costs at the initiation of TRACE. Bessembinder, Maxwell, and Venkataraman [7] find that sophisticated institutional investors benefited from an immediate reduction in trading costs of about $1 billion.

Bessembinder, Maxwell, and Venkataraman [7] estimate their trade execution costs for a sample of institutional (insurance company) trades in corporate bonds before and after the initiation of public transaction reporting for some bonds through the TRACE system in July 2002. They find that the average reduction in one-way trading costs or bonds eligible for TRACE transaction reporting is about 5 to 8 bps. This translates to a reduction in trade execution costs of about 50%. Moreover, they find a 20% reduction for bonds not eligible for TRACE reporting. Bessembinder, Maxwell, and Venkataraman [7] interpret their results as suggesting that better pricing information regarding some bonds also improves valuation and execution cost monitoring for related bonds. They find no evidence that market quality deteriorated in other dimensions.

Bessembinder, Maxwell, and Venkataraman [7] also find that larger trading cost reductions for less liquid and lower-rated bonds, and for larger trades. They estimate that their results equate to annual trading cost reductions of roughly $1 billion per year for the entire corporate bond market, reinforcing that market design can have first-order effects, even for relatively sophisticated institutional customers.

Goldstein, Hotchkiss, and Sirri [20] design and construct a controlled experiment to examine the impact of introducing price transparency on liquidity for BBB-rated corporate bonds. They selected the 120 BBB-rated bonds for which the NASD began disseminating trade data on April 14, 2003. They simultaneously selected a control sample of non-disseminated bonds.

Goldstein, Hotchkiss, and Sirri [20] find that DRT spreads decrease for most BBB-rated corporate bonds whose prices become transparent, and that this effect is strongest for intermediate trade sizes. The only caveat to this result is that they do not find any significant transparency effect for the most thinly-traded bonds. Overall, Goldstein, Hotchkiss, and Sirri [20] conclude that their finds indicate that the introduction of post-trade price transparency has a neutral or positive effect on market liquidity.

The similar results and conclusions in the three complementary TRACE studies collectively generate important policy implications. Foremost, policymakers should take comfort in the fact that there are few, if any, instances in the combined results that show any harm to investors from introducing price transparency to securities markets. To the contrary, the results show that both retail and institutional investors benefit from price transparency. The empirical results from the TRACE studies support the well-founded economic theoretical arguments that transparency should lower transaction costs, especially for smaller trades.

Speeches and testimony by US bond market regulators, such as those listed in the bibliography, show that these studies critically informed the debate over adding price transparency to the US bond markets. Moreover, they continue to provide important lessons for policy makers in bond markets outside of the United States. The bibliography also contains a partial listing of international reports, discussion papers, and conference proceedings that prominently cite the TRACE studies.

Edwards, Harris, and Piwowar (2007)

Edwards et al. [17] apply the Harris and Piwowar [26] econometric approach to corporate bonds. They also extend the approach by allowing liquidity to be time varying. This extension allows them to examine how the introduction of price transparency affects corporate bond transaction costs.

They model the unobserved value return r_{ts}^{V} by decomposing it into the linear sum of a time drift, an average bond index return, differences between index returns for long and short term bonds and for high and low quality bonds, and a bond-specific valuation factor, ε_{ts}.

$$\begin{aligned} r_{ts}^{\mathrm{V}} = {} & \mathit{Days}_{ts}\left(\mathit{DriftRate}\right) + \beta_1 \mathit{AveIndexRet}_{ts} \\ & + \beta_2 \mathit{DurationDif}_{ts} + \beta_3 \mathit{CreditDif}_{ts} + \varepsilon_{ts} \,, \end{aligned}$$

where Days_{ts} counts the number of calendar days between trades t and s, *DriftRate* is the bond coupon rate subtracted from five percent, $\mathit{AveIndexRet}_{ts}$ is the index return for the average bond between trades t and s and $\mathit{DurationDif}_{ts}$ and $\mathit{CreditDif}_{ts}$ are the corresponding differences between index returns for long and short term bonds and high and low credit risk bonds. The first term accounts for the continuously compounded bond price return that traders expect when interest rates are constant and the bond's coupon interest rate differs from five percent. The three factor returns account for bond value changes due to shifts in interest rates and credit spreads. Edwards et al. [17] estimate the bond indices using repeat sale regression methods with terms that account for bond transaction costs. Finally, the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma^2_{\varepsilon_{ts}} = N^{\text{Sessions}}_{ts}\sigma^2_{\text{Sessions}} \,,$$

where N^{Sessions}_{ts} is the number of trading sessions and fractions of trading sessions between trades t and s.

Edwards et al. [17] model customer transaction costs using the following additive expression:

$$c(S_t) = c_0 + c_1\frac{1}{S_t} + c_2 \log S_t + c_3 S_t + c_4 S^2_t + \kappa_t \,,$$

where κ_t represents variation in the actual customer transaction cost that is unexplained by the average transaction cost function. This variation may be random or due to an inability of the average transaction cost function to represent average trade costs for all trade sizes. They assume κ_t has zero mean and variance given by σ^2_κ.

The first three terms of the cost function are the same as in Harris and Piwowar [26], where the constant term allows total transaction costs to grow in proportion to size, the second term characterizes any fixed costs per trade, and the third term allows for costs per bond to vary by trade size. The two additional terms allow more flexibility in the costs to vary by size. Because corporate bonds trade more frequently than municipal bonds, Edwards et al. [17] did not need to be as concerned about degrees of freedom as Harris and Piwowar [26].

Combining the last three equations produces the Edwards et al. [17] version of the Harris and Piwowar [26] transaction cost estimation model:

$$\begin{aligned} r^{\text{P}}_{ts} - \mathit{Days}_{ts}\left(\mathit{DriftRate}\right) &= c_0\left(Q_t - Q_s\right) \\ &+ c_1\left(Q_t\frac{1}{S_t} - Q_s\frac{1}{S_s}\right) + c_2\left(Q_t \log S_t - Q_s \log S_s\right) \\ &+ c_3\left(Q_t S_t - Q_s S_s\right) + c_4\left(Q_t S^2_t - Q_s S^2_s\right) \\ &+ \beta_1 \mathit{AveIndexRet}_{ts} + \beta_2 \mathit{DurationDif}_{ts} + \beta_3 \mathit{CreditDif}_{ts} \\ &+ \eta_{ts} \,, \end{aligned}$$

where the left hand side is simply the continuously compounded bond return expressed as the equivalent rate on a notional five percent coupon bond. Edwards et al. [17] estimate their time-series model in the same way as Harris and Piwowar [26].

Edwards et al. [17] extend Harris and Piwowar [26] by introducing a pooled time-series regression model that they use to estimate average transaction costs for each day for a class of bonds. With this model, they are able to estimate the daily average transaction costs for bonds that became transparent in 2003, and compare these estimates to those for comparable bonds that were either TRACE-transparent throughout 2003 or never TRACE-transparent in 2003.

The pooled time-series regression model that Edwards et al. [17] use to estimate daily transaction costs differs in two respects from the time-series regression model that they use to estimate average transaction costs for a given bond. First, they specify separate average transaction cost functions, $c_T\left(S_t\right)$, for each day T in the sample. Second, to minimize the total number of parameters to be estimated, they use the three-parameter average cost function:

$$c_T\left(S_t\right) = c_{0T} + c_{1T}\frac{1}{S_t} + c_{2T}\log S_t + \kappa_t \,.$$

For a given bond, the change in value between bond trades is modeled as:

$$\log V_t - \log V_s = f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + e_{st} \,,$$

where S is the day on which trade s took place and T is the day on which a subsequent trade t took place, r_J is the common index return (to be estimated) for day J and f_s and f_t, respectively, are the fractions of the S and T trading days overlapped by the period spanned by transactions s and t. This portion of the specification is the same as appears in many paired trade regression index estimation procedures. With these changes, the regression model is

$$\begin{aligned} r^{\text{P}}_{ts} - \mathit{Days}_{ts}\left(5\% - \mathit{CouponRate}\right) &= c_{0T}Q_t - c_{0S}Q_s \\ + c_{1T}Q_t\frac{1}{S_t} - c_{1S}Q_s\frac{1}{S_s} &+ c_{2T}Q_t\log S_t - c_{2S}Q_s\log S_s \\ &+ f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + \eta_{ts} \,. \end{aligned}$$

They use iterated weighted least squares methods, where the weights are equal to the predicted values of the regression of the squared residuals on the independent variables appearing in the residual variation expression. Edwards et al. [17] estimate the model using a three-month wide sliding window that moves forward one month at a time. The time series of coefficient estimates are assembled from the center months of each of the sliding regressions. They compute transaction costs for various transaction sizes by evaluating the estimated transaction cost functions at the given transaction sizes. Using the estimated variance-covariance matrix of the estimators, they also compute daily standard errors of the various daily transaction cost estimates.

Bessembinder, Maxwell, and Venkataraman (2006)

Bessembinder, Maxwell, and Venkataraman [7] develop and estimate an indicator variable regression model:

$$\Delta P = a + wX_t + \gamma SQ_t^* + \alpha S\Delta Q + \omega_t \,,$$

where ΔP is the change in the price of the bond from time $t-1$ to time t, a is the regression intercept, w is the fraction of public information that is observable in the data with realizations X_t, γS is the informational component of the spread, αS is the non-informational component of the spread (where $\alpha = (1-\gamma)$), Q_t^* is the market maker's estimate of bond value due to surprises in order flow, ΔQ is the change in indicator variable Q (which takes a value of 1 if the trade is a customer buy and -1 if it is a customer sell) from time $t-1$ to time t, and ω_t is the regression error term.

Bessembinder, Maxwell, and Venkataraman [7] develop this regression model in the following way. $E_t(V)$ is the market-maker's estimate of the bond's unobserved true value (V) at time t conditional on whether the observed trade is a customer buy or a customer sell. Transaction prices are given by:

$$P_t = E_t\,(V) + \alpha SQ_t \,.$$

The market maker's estimate of bond value at time t, $E_t(V)$, is her estimate from the prior period, $E_{t-1}(V)$, updated to reflect surprises in order flow, $Q_t - E_{t-1}\,(Q_t)$, and public information revealed since the prior period, η_τ Substitution yields:

$$E_t\,(V) = E_{t-1}\,(V) + \gamma SQ_t^* + \eta_t \,,$$

where $Q_t^* = Q_t - E_{t-1}\,(Q_t)$. To allow for the possibility that bond market order flow is positively autocorrelated, Bessembinder, Maxwell, and Venkataraman [7] assume that it follows a simple AR1 process, so that $E_{t-1}\,(Q_t) = \rho\,(Q_{t-1})$. The change in the price of the bond from time $t-1$ to time t is:

$$P_t - P_{t-1} = \gamma SQ_t^* + \alpha SQ_t - \alpha SQ_{t-1} + \eta_t \,.$$

Substituting ΔP for P_t – P_{t-1} and ΔQ for Q_t – Q_{t-1}, this expression can be rewritten as:

$$\Delta P = \gamma SQ_t^* + \alpha S\Delta Q + \eta_t \,.$$

To incorporate observable public information that affects bond value, they assume that a fraction w of public information becomes observable in the data with realizations X_t, while the remaining portion $(1-w)$ is due to unobservable innovations U_t that represent statistical noise. Substitution yields their regression model:

$$\Delta P = wX_t + \gamma SQ_t^* + \alpha S\Delta Q + \omega_t \,,$$

where $\omega = (1-w)\,U_t$. Bessembinder, Maxwell, and Venkataraman [7] show that their model is equivalent to the Madhavan et al. [39] model. Moreover, in the special case of no autocorrelation in order flow ($\rho = 0$), their model is equivalent to Huang and Stoll [33], Schultz [42] and Warga [44].

Goldstein, Hotchkiss, and Sirri (2007)

Goldstein, Hotchkiss, and Sirri [20] use two different methods to estimate transaction costs for a sample of BBB-rated bonds. Their first method involves identifying "dealer round-trip" (DRT) transaction chains. These transaction chains involve a dealer purchasing a bond from a customer and then selling that same bond to another customer within a specified period of time. DRT spreads are calculated as the difference between the customer buy price at the end of the transaction chain and the customer sell price at the beginning of the chain. Their DRT method is similar to the methods used in the municipal bond studies of Green et al. [22] and Biais and Green [9], except their data contains individual dealer identifiers. Goldstein, Hotchkiss, and Sirri [20] estimate DRT spreads for transaction chains that occur with one-day, five-day, and unlimited time intervals. They estimate DRT spreads for various trade size groups.

Their second method is a regression method similar to Warga [44] and Schultz [42]. For each trade size group, Goldstein, Hotchkiss, and Sirri [20] estimate spreads by regressing the difference between the transaction price for a customer (T) and an estimated bid price (B) on a dummy variable that equals one for customer buys and zero for customer sells:

$$[T-B]_i = \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \varepsilon_i \,,$$

where estimated bid prices are obtained from Reuters dealer bid price estimates from the end of the day prior to the transaction. Reuters estimates are based on daily quotes obtained directly from individual dealers.

Goldstein, Hotchkiss, and Sirri [20] estimate a second regression to consider the effect of dissemination while controlling for other bond characteristics impacting the spread:

$$[T-B]_i = \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \alpha_2 D_i^{\text{DisseminatedBond}} + \alpha_3 D_i^{\text{Post-disseminationPeriod}} + \alpha_4 D_i^{\text{DisseminationBond* Post-disseminationPeriod}} + \alpha_5 X_5 + \cdots + \alpha_{10} X_{10} + \varepsilon_i \, ,$$

where additional dummies are included for disseminated bonds, transactions that occur in the post-dissemination period, and the interaction of these two dummies for transaction in disseminated bonds that occur in the post-dissemination period. As in Schultz [42] the additional dummies are expressed a $+1$ for buys and -1 for sells. Variables $X_5, \ldots, X_{10}$ are six bond characteristics related to spreads: trade size, time-to-maturity, age, issue amount, average daily volume over the prior 30 days, and days since last trade.

The Links Between Bond Market Microstructure Research and Other Finance and Economics Research

The discussion of bond market research has thus far been presented solely within the framework of the market microstructure literature. However, some of the most interesting bond market research is connected to other areas of finance and economics. The instrument complexity results of Harris and Piwowar [26], for example, provide evidence to support Carlin's [11] formal model of strategic price complexity in security design for retail markets. Additionally, Chen, Lesmond, and Wei [13] provide an example of how bond market microstructure research is linked to asset pricing models in finance. Green, Hollifield, and Schurhoff [23] develop a theoretical model that is analogous to the costly consumer search models in the broader economics literature.

Chen, Lesmond, and Wei (2007)

Beginning with Amihud and Mendelson [3], market microstructure research has consistently shown that a liquidity premium exists in equity markets. Recently, bond market microstructure researchers have begun investigating whether a liquidity premium also exists in bond markets. One such paper is Chen, Lesmond, and Wei [13]. Their investigation of the link between bond liquidity and corporate yield spreads provides important implications for the default risk literature.

Chen, Lesmond, and Wei [13] investigate whether liquidity is priced in corporate yield spreads. They use several approaches, including a regression approach that is an extension to the Lesmond, Ogden, and Trzcinka [36] (LOT) approach developed for equities. The LOT approach assumes that a zero return day (or a non-trading day) is observed when the true price changes by less than the transaction costs. Because marginal informed investors will only trade on information if the trade yields expected profits net of transaction costs, an individual bond's trading costs represent a threshold that must be exceeded before its return will reflect new information. The premise of this approach is that if the value of the information is insufficient to exceed the costs of trading, then the marginal investor will either reduce trading or not trade, causing a zero return.

Chen, Lesmond, and Wei [13] extend the LOT approach to corporate bonds by applying a two-factor return-generating model to estimate bond trading costs:

$$R^*_{j,t} = \beta_{j1} \mathit{Duration}^*_{j,t} \Delta R_{ft} + \beta_{j2} \mathit{Duration}^*_{j,t} \Delta \mathit{S\&P\ Index}_t + \varepsilon_{j,t} \, ,$$

where $R^*_{j,t}$ is the unobserved "true" return for bond j on day t that investors would bid given zero trading costs. The daily change in the 10-year risk-free interest rate, ΔR_{ft}, is the factor that is more important for investment grade bonds, while the second factor, $\Delta \mathit{S\&P\ Index}_t$, the daily return on the Standard & Poor's 500 equity index, is more important for high-yield bonds. Both factors are scaled by $\mathit{Duration}_{j,t}$, the bond's duration.

Chen, Lesmond, and Wei [13] then apply the Amihud and Mendelson [3] liquidity premium to bonds. In Amihud and Mendelson [3], observed asset prices differ from true values because of a liquidity premium that compensates investors for liquidity costs. Chen, Lesmond, and Wei [13] state the liquidity effects on bond returns as:

$$R_{j,t} = R^*_{j,t} - \alpha_{i,j} \, ,$$

where $R_{j,t}$ is the measured return, $\alpha_{2,j}$ is the effective buy-side cost, and $\alpha_{1,j}$ is the effective sell-side cost for bond j. The resulting liquidity constraint is:

$$\begin{array}{lll} R_{j,t} = R^*_{j,t} - \alpha_{1j} & \text{if } R^*_{j,t} < \alpha_{1j} & \text{and } \alpha_{1j} < 0 \\ R_{j,t} = 0 & \text{if } \alpha_{1j} \le R^*_{j,t} \le \alpha_{2j} & \\ R_{j,t} = R^*_{j,t} - \alpha_{2j} & \text{if } R^*_{j,t} > \alpha_{2j} & \text{and } \alpha_{2j} > 0 \, . \end{array}$$

Combining the liquidity constraint with the return generating model, Chen, Lesmond, and Wei [13] us a maximum likelihood method outlined in LOT to estimate transaction costs. They specify the log-likelihood function as:

$$
\begin{aligned}
\mathrm{Ln}L = {} & \sum_{1} \mathrm{Ln} \frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\
& - \sum_{1} \frac{1}{2\sigma_j^2} \left(R_j + \alpha_{1,j} - \beta_{j1} Duration_{j,t} * \Delta R_{ft}\right. \\
& \qquad \left. - \beta_{j2} Duration_{j,t} * \Delta S\&P\ Index\right)^2 \\
& + \sum_{2} \mathrm{Ln} \frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\
& + \sum_{2} \frac{1}{2\sigma_j^2} \left(R_j + \alpha_{2,j} - \beta_{j1} Duration_{j,t} * \Delta R_{ft}\right. \\
& \qquad \left. - \beta_{j2} Duration_{j,t} * \Delta S\&P\ Index\right)^2 \\
& + \sum_{0} \mathrm{Ln} \left(\Phi_{2,j} - \Phi_{1,j}\right) ,
\end{aligned}
$$

where $\Phi_{i,j}$ represents the cumulative distribution function for each bond-year evaluated at:

$$
\frac{\left(\alpha_{1,j} - \beta_{j1} Duration_{j,t} * \Delta R_{ft}\right)}{\sigma_j} - \frac{\left(\beta_{j2} Duration_{j,t} * \Delta S\&P\ Index_t\right)}{\sigma_j} .
$$

Σ_1 (region 1) represents the negative nonzero measured returns, Σ_2 (region 2) represents the positive nonzero measured returns, and Σ_0 (region 0) represents the zero measured returns. The difference in the buy-side and sell-side cost estimates, $\alpha_{2,j} - \alpha_{1,j}$, represents round-trip trading costs.

The model's implicit assumption that information motivates bond trades and that information is efficiently impounded into bond prices is supported by the results of Hotchkiss and Ronen [31]. The error term captures noise trading and trades due to unanticipated public information.

In addition to LOT estimates, Chen, Lesmond, and Wei [13] use bid-ask spreads and zero-returns as liquidity cost measures. The bid-ask spreads the use are bond-year proportional bid-ask spreads, calculated as the average of quarterly proportional spreads. Quarterly proportional spreads are calculated from quarterly bid-ask spreads obtained from Bloomberg consensus quotes among market participants, divided by the average bid and ask price. Zero-returns are simply the percentage of days with returns equal to zero.

They find that liquidity costs are related to credit rating. Liquidity costs are much higher for high-yield bonds than for investment grade bonds. They also find that liquidity costs are related to maturity. Liquidity costs for long-maturity bonds are higher than for short-maturity bonds.

They also find that yield spreads generally increase with maturity for investment grade bonds. But, they find that yield spreads generally decrease with maturity for high-yield bonds. They point out the endogeneity issue stemming from the Helwege and Turner [28] finding that relatively safer firms within the same high-yield credit rating category tend to issue longer-term bonds. This endogeneity issue causes the average yield spread to decline with maturity for high-yield bonds.

To investigate whether liquidity is priced in corporate yield spreads, Chen, Lesmond, and Wei [13] first run the following regression specification for investment grade bonds and high-yield bonds separately:

$$
\begin{aligned}
Yield\ Spread_{it} = {} & \eta_0 + \eta_1 Liquidity_{it} + \eta_2 Maturity_{it} \\
& + \eta_3 Amount\ Outstanding_{it} + \eta_4 Coupon_{it} \\
& + \eta_5 Treasury\ Rate_{it} + \eta_6 10Yr - 2Yr\ Treasury\ Rate_{it} \\
& + \eta_7 EuroDollar_{it} + \eta_8 Volatility_{it} + \eta_9 Bond\ Rating_{it} \\
& + \eta_{10} PreTax\ Coverage\ Dummy_{it} \\
& + \eta_{11} Operating\ Income/Sales_{it} + \eta_{12} Debt/Assets_{it} \\
& + \eta_{13} Debt/Capitalization_{it} + \varepsilon_{it} ,
\end{aligned}
$$

where the subscript *it* refers to bond *i* in year *t*. *Liquidity* refers to the three liquidity cost measures – bid-ask spread, zero-returns, or the LOT estimate. Additional variables control for bond-specific, firm-specific, and macroeconomic factors. *Maturity* is the number of years until the bond matures relative to the year being analyzed, and *Amount Outstanding* is natural logarithm of the dollar amount outstanding, *Coupon* is the bond coupon rate. *Treasury Rate* is the 1-year Treasury Note rate, *10Yr-2Yr Treasury Rate* is the difference between the 10-year and 2-year Treasury rates, and *Eurodollar* is the 30-day Eurodollar rate minus the 3-month T-Bill rate. *Volatility* is the equity volatility for each issuer and *Bond Rating* is a credit rating scale that ranges from 1 (AAA rating) to 10 (BBB- rating) for investment grade bonds and from 1(BB+ rating) to 12 (D rating) for high-yield bonds. *PreTax Coverage Dummy* represents four dummy variables corresponding to groupings of pre-tax income, *Operating Income/Sales*, *Debt/Assets*, and *Debt/Capitalization* are each firm's respective accounting ratios.

They find that all three liquidity measures are positively related to the yield spread in both the investment grade and high-yield samples. The liquidity coefficients are statistically significant at the 1% level in every scenario. This provides strong evidence that liquidity is priced in corporate yield spreads. This finding is robust to controlling for issuer influences with issuer fixed-effects regressions. The only caveat is that they achieve slightly weaker results for the zero-return liquidity cost measure than for bid-ask spreads and LOT estimates. This finding is also robust to controlling for potential endogeneity problems arising from the contemporaneous measurement of the yield spread, liquidity costs, and credit rating. They perform this robustness check by employing a simultaneous equations model using three equations that correspond to each of the potentially endogenous variables:

$$\begin{aligned}
&\mathit{Yield\ Spread}_{it} = \eta_0 + \eta_1 \mathit{Liquidity}_{it} + \eta_2 \mathit{Maturity}_{it}\\
&+ \eta_3 \mathit{Coupon}_{it} + \eta_4 \mathit{Treasury\ Rate}_{it}\\
&+ \eta_5 10\mathit{Yr} - 2\mathit{Yr\ Treasury\ Rate}_{it} + \eta_6 \mathit{EuroDollar}_{it}\\
&+ \eta_7 \mathit{Volatility}_{it} + \eta_8 \mathit{Bond\ Rating}_{it}\\
&+ \eta_9 \mathit{PreTax\ Coverage\ Dummy}_{it}\\
&+ \eta_{10} \mathit{Operating\ Income/Sales}_{it} + \eta_{11} \mathit{Debt/Assets}_{it}\\
&\qquad\qquad + \eta_{12} \mathit{Debt/Capitalization}_{it} + \varepsilon_{it}\ ,
\end{aligned}$$

$$\begin{aligned}
&\mathit{Liquidity}_{it} = \eta_0 + \eta_1 \mathit{Maturity}_{it} + \eta_2 \mathit{Age}_{it}\\
&+ \eta_3 \mathit{Amount\ Outstanding}_{it} + \eta_4 \mathit{Bond\ Rating}_{it}\\
&\quad + \eta_5 \mathit{Bond\ Volatility}_{it} + \eta_6 \mathit{Yield\ Spread}_{it} + \varepsilon_{it}\ ,
\end{aligned}$$

$$\begin{aligned}
&\mathit{Credit\ Rating}_{it} = \eta_0 + \eta_1 \mathit{Treasury\ Rate}_{it}\\
&+ \eta_2 10\mathit{Yr} - 2\mathit{Yr\ Treasury\ Rate}_{it}\\
&+ \eta_3 \mathit{PreTax\ Coverage\ Dummy}_{it}\\
&+ \eta_4 \mathit{Operating\ Income/Sales}_{it}\\
&+ \eta_5 \mathit{Debt/Assets}_{it} + \eta_6 \mathit{Debt/Capitalization}_{it}\\
&\qquad\qquad + \eta_7 \mathit{Yield\ Sprad}_{it} + \varepsilon\ .
\end{aligned}$$

The model is estimated using twostage least squares. The simultaneous equation model estimation results show that the potential endogeneity does not affect the relation between liquidity and yield spreads for either the investment grade or the high-yield bonds.

Thus, Chen, Lesmond, and Wei [13] find extremely consistent and robust evidence that liquidity is a key determinant in corporate yield spreads. This finding provides at least a partial explanation for the findings of Collin-Dufresne, Goldstein, and Martin [15] and others who show that default risk does not completely explain corporate yield spreads.

Green, Hollifield, and Schurhoff (2007b)

Green, Hollifield, and Schurhoff [23] examine secondary market trading in newly issued municipal bonds for the first 60 trading days of their lives. They begin by descriptively documenting the price behavior of newly issued municipal bonds. They show that municipal bonds are underpriced when issued. But, unlike equities, the average price rises slowly over a period of several days. Green, Hollifield, and Schurhoff [23] also find that the observed price patterns are complex. High levels of price dispersion are observed for small trade sizes in the aftermarket for new municipal bond issues. While some small traders purchase bonds on attractive terms, others do not. In contrast, there is very little price dispersion for large trade sizes. Virtually all of the large traders purchase bonds on attractive terms.

They argue that the price level and dispersion patterns are the result of bond dealers discriminating between informed and uninformed customers. Accordingly, Green, Hollifield, and Schurhoff [23] develop and estimate a mixed distribution model for the markups that uninformed and informed investors pay when they purchase newly issued bonds. Their model incorporates investor search costs, i. e., the costs in terms of time and effort needed for investors to become informed about new bond issues.

The mixed distribution model of Green, Hollifield, and Schurhoff [23] is analogous to economic models of costly consumer search, such as the gametheoretic model of Shilony [43] that focuses on advertising and price competition among retail stores in homogeneous product markets. Shilony [43] assumes that all stores are required to advertise, but the advertising is segmented (e. g., signs are posted on store windows). Consumers have a preference for the particular store that that offers them free access to the advertising (e. g., the store right outside their house or the one that they regularly visit) and they will pay more for a product at this store even if it does not offer the lowest price.

The institutional mechanisms of the primary market and the structure of the secondary market for municipal bonds fits particularly well with the informational interpretation of Shilony's [43] model. Every investor has free information about the price that will be charged by his broker. Also, because all firms must disseminate their last sale information on a real-time basis, the investor can choose to look on www.investinginbonds.com or some other free website to find the range of prices charged by all brokers. But, this last sale information does not identify which broker charged the lowest price. To find this out, the investor must incur some cost.

Green, Hollifield, and Schurhoff [23] begin with the assumption that there are both observable and unobservable sources of heterogeneity in the costs investors face in gathering and using information about prices of new municipal issues. They assume that for investor i, the difference between the benefit and the cost of learning about a new issue is z_i^* with:

$$z_i^* = w_i\delta + \mu_i \ ,$$

where w_i is a vector of conditioning variables, δ is a parameter vector, and μ_ι is an error term. The error term is observed by the investor, but not by the econometrician. Investor i becomes informed about the price of a new issue if and only if $z_i^* \geq 0$. They do not observe z_i^*, but they do observe w_i and the price the investor pays for the bond.

An investor who is uninformed about the reoffering price for a new bond is willing to pay the percentage markup y_U of:

$$y_{Ui} = x_i\beta + \varepsilon_{Ui} \ ,$$

where x_i is a vector of conditioning variables, β is a parameter vector, and ε_{Ui} is an error term. Similarly, an investor who is informed about the underwriter's pricing of a new bond is willing to pay the percentage markup y_I of:

$$y_{Ii} = x_i\gamma + \varepsilon_{Ii} \ ,$$

where x_i is a vector of conditioning variables, γ is a parameter vector, and ε_{Ii} is an error term. The uncertainty about the percentage markup is expected to be lower when the investor is informed than when the investor is uninformed:

$$\sigma_I < \sigma_U \ .$$

They use this condition to empirically identify the informed versus uninformed distributions from which the observed markups, y_i, are drawn:

$$y_i = \begin{cases} y_{Ui} & \text{if} \quad z_i^* < 0 \,, \\ y_{Ii} & \text{if} \quad z_i^* \geq 0 \,. \end{cases}$$

Green, Hollifield, and Schurhoff [23] use iterated expectations to show that investors take the markup into account when deciding whether to become informed about an upcoming bond issue or not:

$$\begin{aligned} E\left(y_i \left| w_i, x_i \right.\right) = E\left(y_i \left| \text{Informed}_i, w_i, x_i \right.\right) \\ \Pr\left(\text{Informed}_i \left| w_i \right.\right) \\ + E\left(y_i \left| \text{Uninformed}_i, w_i, x_i \right.\right) \Pr\left(\text{Uninformed}_i \left| w_i \right.\right) . \end{aligned}$$

They estimate their model under the assumption that the error terms are drawn independently and identically from a multivariate normal distribution:

$$\begin{pmatrix} u_i \\ \varepsilon_{Ui} \\ \varepsilon_{Ii} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_U\sigma_U & \rho_I\sigma_I \\ \rho_U\sigma_U & \sigma_U^2 & 0 \\ \rho_I\sigma_I & 0 & \sigma_I^2 \end{bmatrix}\right),$$

where ρ_U is the correlation between u_i and ε_{Ui} and ρ_I is the correlation between u_i and ε_{Ii}. Denoting the cumulative standard normal distribution as Φ and the standard normal density as φ, Green, Hollifield, and Schurhoff [23] show that the condition that investor i becomes informed if and only if $z_i^* \geq 0$ implies that:

$$\begin{aligned} \Pr\left(\text{Informed}_i \left| w_i \right.\right) &= \Pr\left(z_i^* \geq 0 \left| w_i \right.\right) \\ &= \Pr\left(u_i \geq -w_i\delta \left| w_i \right.\right) \\ &= \Phi\left(w_i\delta\right) \ , \end{aligned}$$

and

$$\Pr\left(\text{Uninformed}_i \left| w_i \right.\right) = 1 - \Phi\left(w_i\delta\right) \ .$$

By combining equations and using the distributional assumptions of the error terms, Green, Hollifield, and Schurhoff [23] show that

$$E\left(y_i \left| \text{Informed}_i, w_i, x_i \right.\right) = x_i\gamma + \rho_I\sigma_I\frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)} \ ,$$

and

$$E\left(y_i \left| \text{Uninformed}_i, w_i, x_i \right.\right) = x_i\beta + \rho_U\sigma_U\frac{-\phi\left(w_i\delta\right)}{1 - \Phi\left(w_i\delta\right)} .$$

Therefore, the expected markup is:

$$\begin{aligned} E\left(y_i \left| w_i, x_i \right.\right) = \left(x_i\gamma + \rho_I\sigma_I\frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)}\right)\Phi\left(w_i\delta\right) \\ + \left(x_i\beta + \rho_U\sigma_U\frac{-\phi\left(w_i\delta\right)}{1 - \Phi\left(w_i\delta\right)}\right)1 - \Phi\left(w_i\delta\right) \ . \end{aligned}$$

Green, Hollifield, and Schurhoff [23] estimate this equation as a switching regression. Their results are consistent with two pricing regimes for newly issued municipal bonds. Uninformed investors pay higher average prices than informed investors and there is very little variation in prices paid by informed investors. They also find that the upward trend in prices after issuance is related to the change in the mix if informed and uninformed investors. Informed investors buy soon after issuance, while uninformed investors buy later on.

With respect to the decision about whether to become informed about a new municipal bond issue, they find that large buyers are more likely to become informed than small buyers. To examine how much money is left on the table by uninformed investors who pay large markups to dealers, Green, Hollifield, and Schurhoff [23] classify each transaction into either the Informed or Uninformed regime. The classification is based on whether the expected benefit from becoming informed about a new bond issue is greater than the cost of doing so:

$$\begin{aligned}\text{Informed}_i = 1 &\Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) \geq 0,\\ \text{Uninformed}_i = 1 &\Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) < 0\,.\end{aligned}$$

The difference in the expected markup between an informed investor and an uninformed investor is:

$$\begin{aligned}E\left(y_{Ui} y_{Ii} \,|\text{Uninformed}_i, w_i, x_i\right) &= x_i\left(\beta - \gamma\right)\\ &\quad + \left(\rho_U \sigma_U - \rho_I \sigma_I\right) \frac{-\phi\left(w_i \delta\right)}{1 - \Phi\left(w_i \delta\right)}\,.\end{aligned}$$

They define the money left on the table in each transaction with an uninformed investor as

$$\Delta_i = \begin{cases} \max\left\{E\left(y_{Ui} y_{Ii} \,|\text{Uninformed}_i, w_i, x_i\right), 0\right\}, & \\ & \text{if Uninformed}_i = 1,\\ 0, & \text{else}\,.\end{cases}$$

They denote the estimates of Δ_i as $\widehat{\Delta}_i$ and obtain a cumulative measure across all sales transactions i in a given bond issue j, and then across all issues in a bond deal:

$$\widehat{\text{Money Left on the Table}} = \sum_{\text{Issues}\,j} \sum_{i \in j} \widehat{\Delta}_i\,.$$

Green, Hollifield, and Schurhoff [23] find that money left on the table by uninformed investors represents a significant fraction of the overall expected profits to the underwriters and dealers.

Bibliography

Primary Literature

1. Alexander G, Edwards A, Ferri M (2000a) The determinants of trading volume of high-yield corporate bonds. J Financ Mark 3:177–204
2. Alexander G, Edwards A, Ferri M (2000b) What does Nasdaq's high-yield bond market reveal about bondholderstockholder conflicts? Financ Manag 29:23–39
3. Amihud Y, Mendelson H (1986) Asset pricing and the bid-ask spread. J Financ Econ 17:223–249
4. Amihud Y, Mendelson H (1991) Liquidity, maturity, and yields on US Treasury securities. J Financ 46:1411–1425
5. Barclay M, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from Treasuries going off the run. J Financ 61:2395–2414
6. Bernhardt D, Dvoracek V, Hughson E, Werner I (2005) Why do larger orders receive discounts on the London Stock Exchange? Rev Financ Stud 18:1343–1368
7. Bessembinder H, Maxwell W, Venkataraman K (2006) Optimal market transparency: Evidence from the initiation of trade reporting in corporate bonds. J Financ Econ 82:251–288
8. Biais B, Glosten L, Spatt C (2005) Market microstructure: A survey of microfoundations, empirical results, and policy implications. J Financ Mark 8:217–264
9. Biais B, Green R (2005) The microstructure of the bond market in the 20th century. Working paper, Carnegie Mellon University
10. Blume M, Keim D, Patel S (1991) Returns and volatility of low-grade bonds. J Financ 46:49–74
11. Carlin B (2007) Strategic price complexity in retail financial markets. Working paper, UCLA
12. Chakravarty S, Sarkar A (2003) Trading costs in three US bond markets. J Fixed Income 13:39–48
13. Chen L, Lesmond D, Wei J (2007) Corporate yield spreads and bond liquidity. J Financ 52:119–149
14. Cornell B, Green K (1991) The investment performance of low-grade bond funds. J Financ 46:29–48
15. Collin-Dufresne P, Goldstein R, Martin S (2001) The determinants of credit spread changes. J Financ 41:2177–2207
16. Downing C, Zhang F (2004) Trading activity and price volatility in the municipal bond market. J Financ 59:899–931
17. Edwards A, Harris L, Piwowar M (2007) Corporate Bond Market Transaction Costs and Transparency. J Financ 62:1421–1451
18. Elton E, Green C (1998) Tax and liquidity effects in pricing government bonds. J Financ 53:1533–1562
19. Fenn G (2000) Speed of issuance and the adequacy of disclosure in the 144a high-yield debt market. J Financ Econ 56:383–405
20. Goldstein M, Hotchkiss E, Sirri E (2007) Transparency and liquidity: A controlled experiment on corporate bonds. Rev Financ Stud 20:235–273
21. Green R (2007) Issuers, underwriter syndicates, and aftermarket transparency. Presidential Address to the American Finance Association
22. Green R, Hollifield B, Schurhoff N (2007a) Financial intermediation and the costs of trading in an opaque market. Rev Financ Stud 20:275–314
23. Green R, Hollifield B, Schurhoff N (2007b) Dealer intermediation and price behavior in the aftermarket for new bond issues. J Financ Econ 86:643–682
24. Harris L (2003) Trading and exchanges: market microstructure for practitioners. Oxford University Press, New York
25. Harris L, Piwowar M (2004) Municipal bond liquidity. Working paper, Available at SSRN: http://ssrn.com/abstract=503062
26. Harris L, Piwowar M (2006) Secondary trading costs in the municipal bond market. J Financ 61:1361–1397
27. Hasbrouck J (1993) Assessing the quality of a security market: A new approach to transaction-cost measurement. Rev Financ Stud 6:191–212
28. Helwege J, Turner C (1999) The slope of the credit yield curve for speculative-grade issuers. J Financ 54:1869–1884
29. Hong G, Warga A (2000) An empirical study of bond market transactions. Financ Anal J 56:32–46
30. Hong G, Warga A (2004) Municipal marketability. J Fixed Income 14:86–95
31. Hotchkiss E, Ronen T (2002) The informational efficiency of the corporate bond market: An intraday analysis. Rev Financ Stud 15:1325–1354

32. Hotchkiss E, Warga A, Jostova G (2002) Determinants of corporate bond trading: A comprehensive analysis. Working paper, Boston College
33. Huang R, Stoll H (1997) The components of the bid-ask spread: A general approach. R Financ Stud 10:995–1034
34. Kalimipalli M, Warga A (2002) Bid/ask spread, volatility and volume in the corporate bond market. J Fixed Income 12:31–42
35. Karolyi GA (2004) The world of cross-listings and cross-listings of the world: Challenging conventional wisdom. Working paper, Ohio State University
36. Lesmond D, Ogden J, Trzcinka C (1999) A new estimate of transaction costs. Rev Financ Stud 12:1113–1141
37. Levitt A (1998) The importance of transparency in America's debt market. Remarks at the Media Studies Center, New York, September 9, 1998
38. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–208
39. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. R Financ Stud 10:1035–1064
40. Reiss P, Werner I (1996) Transaction costs in dealer markets: Evidence from the London Stock Exchange. In: Lo A (ed) The industrial organization and regulation of the securities industry. University of Chicago Press, Chicago
41. Sarig O, Warga A (1989) Bond price data and bond market liquidity. J Financ Quant Anal 24:367–378
42. Schultz P (2001) Corporate bond trading costs: A peek behind the curtain. J Financ 56:677–698
43. Shilony Y (1977) Mixed pricing in oligopoly. J Econ Theory 14:373–88
44. Warga A (1991) Corporate bond price discrepancies in the dealer and exchange markets. J Fixed Income 1:7–16
45. Warga A (1992) Bond returns, liquidity and missing data. J Financ Quant Anal 27:605–616

Books and Reviews

Greene W (1993) Econometric Analysis, 2nd edn. Macmillan, New York

O'Hara M (1997) Market microstructure theory. Basil Blackwell, Cambridge

Shultz B (1946) The securities market and how it works. Harper, New York

US Securities and Exchange Commission (2004) Report on transactions in municipal securities. http://www.sec.gov/news/studies/munireport2004.pdf

Speeches, Public Statements, and Testimony by US Bond Market Regulators and Participants

September 9, 1998, The Importance of Transparency in America's Debt Market, Remarks of SEC Chairman Arthur Levitt at the Media Studies Center, New York, NY, http://www.sec.gov/news/speech/speecharchive/1998/spch218.htm

April 23, 1999, Remarks of SEC Chairman Arthur Levitt Before the Bond Market Association, San Francisco, CA, http://www.sec.gov/news/speech/speecharchive/1999/spch268.htm

October 28, 1999, Electronic Trading Technology's Impact on the Fixed-Income Markets, Remarks of SEC Commissioner Laura S. Unger at The Bond Market Association, Fifth Annual Legal and Compliance Seminar, New York, NY, http://www.sec.gov/news/speech/speecharchive/1999/spch313.htm

January 8, 2002, Remarks Before the Bond Market Association Legal and Compliance Conference, by Annette L. Nazareth, SEC Director, Division of Market Regulation, http://www.sec.gov/news/speech/spch532.htm

April 25, 2002, Remarks Before the Annual Meeting of the Bond Market Association, by SEC Chairman Harvey L. Pitt, New York, NY, http://www.sec.gov/news/speech/spch553.htm

February 3, 2004, Legal & Compliance Conference – The Bond Market Association, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/spch020304cag.htm

June 17, 2004, US Senate Committee on Banking, Housing, and Urban Affairs, Hearing on "An Overview of the Regulation of the Bond Markets", http://banking.senate.gov/public

September 9, 2004, Remarks before the 2004 Bond Attorney's Workshop of the National Association of Bond Lawyers, by Martha Mahan Haines, SEC Assistant Director, Office of Municipal Securities, Division of Market Regulation, Chicago, IL, http://www.sec.gov/news/speech/spch090904mmh.htm

October 1, 2004, Keynote Address before the NASD Conference on Fixed Income Markets, by Annette L. Nazareth, SEC Director, Division of Market Regulation, New York City, NY, http://www.sec.gov/news/speech/spch100104aln.htm

December 16, 2004, Second MTS Conference on Financial Markets: "The Organization and Performance of Fixed-Income Markets", by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Vienna, Austria, http://www.sec.gov/news/speech/spch121604cs.htm

February 1, 2005, Remarks before the Bond Market Association 2005 Legal & Compliance Conference, by SEC Commissioner Paul S. Atkins, New York, NY, http://www.sec.gov/news/speech/spch020105psa.htm

February 2, 2005, Keynote Address at the Bond Market Association 2005 Legal & Compliance Conference, by NASD Vice Chairman Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013225

April 20, 2005, Remarks before the Bond Market Association, by SEC Chairman William H. Donaldson, New York, NY, http://www.sec.gov/news/speech/spch042005whd.htm

April 20, 2005, Navigating the Changing Climate in Fixed-Income Products, Remarks of NASD President Doug Shulman before the Bond Market Association, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013842

May 6, 2005, Broad Themes in Market Microstructure, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Cambridge, MA, http://www.sec.gov/news/speech/spch050605css.htm

June 21, 2005, Developing Bond Markets in APEC: Key Lessons from the US Experience, Remarks of SEC Commissioner Roel C. Campos before the ABAC/ADBI/PECC Conference, Tokyo, Japan, http://www.sec.gov/news/speech/spch062105rcc-2.htm

September 13, 2005, Remarks for Promethee: Transatlantic Dialogue and Regulatory Convergence: Panel on Financial Markets, by Ethiopis Tafara, SEC Director, Office of International Affairs, Paris, France, http://www.sec.gov/news/speech/spch091305et.htm

September 15, 2005, Keynote Address before the 30th Bond Attorney's Workshop of the National Association of Bond Lawyers,

by SEC Commissioner Roel C. Campos, Chicago, IL, http://www.sec.gov/news/speech/spch091505rcc.htm

November 17, 2005, Address by NASD President Doug Shulman to the NASD Fall Securities Conference, San Francisco, CA, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P015554

January 6, 2006, Discussion: An Overview of Bond Market Transparency, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Boston, MA, http://www.sec.gov/news/speech/spch010606css.htm

February 7, 2006, Remarks before the TBMA Legal and Compliance Conference, by SEC Commissioner Annette L. Nazareth, New York, NY, http://www.sec.gov/news/speech/spch020706aln.htm

May 19, 2006, The Continuing Evolution of the Bond Market and Other Observations: Remarks Before the Bond Market Association's 30th Annual Conference, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/2006/spch051906cag.htm

May 19, 2006, Remarks Before the Bond Market Association's 30th Annual Conference, by NASD President Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P016651

Miscellaneous Sources of Information on International Corporate Bond Markets

April, 2004, Markets in Financial Instruments Directive (MiFID), Directive 2004/39/EC of the European Parliament and of the Council, Article 65, http://europa.eu.int/eur-lex/pri/en/oj/dat/2004/l_145/l_14520040430en00010044.pdf

May 2004, International Organization of Securities Commissions (IOSCO), "Transparency of Corporate Bond Markets", Report of the Technical Committee of IOSCO, http://www.iosco.org/library/pubdocs/pdf/IOSCOPD168.pdf

September 2005, UK Financial Services Authority (FSA), "Trading Transparency in the UK Secondary Bond Markets", FSA Discussion Paper 05/5, http://www.fsa.gov.uk/pubs/discussion/dp05_05.pdf

November, 2005, "Developing Bond Markets in Asia Conference", jointly by the Asian Office of the Bank for International Settlements (BIS) and the People's Bank of China (PBC), Kunming, China, http://www.bis.org/publ/bppdf/bispap26.htm

May 2006, Centre for Economic Policy Research (CEPR), "European corporate bond markets: Transparency, liquidity, efficiency", http://www.cepr.org/PRESS/TT_CorporateFULL.pdf

Delay and Disruption in Complex Projects

SUSAN HOWICK[1], FRAN ACKERMANN[1], COLIN EDEN[1], TERRY WILLIAMS[2]

[1] Strathclyde Business School, University of Strathclyde, Glasgow, UK

[2] School of Management, Southampton University, Southampton, UK

Article Outline

Glossary

Cause map A cause map is similar to a cognitive map however it is not composed of an individuals perception but rather the views/statements from a number of participants. It follows the same formalisms as cognitive mapping but does not reflect cognition as it is composite.

Cognitive map A cognitive map is a representation of an individuals perception (cognition) of an issue. It is graphically depicted illustrating concepts/statements connected together with arrows representing causality. They are created using a set of established formalisms.

Complex project A complex project is a project in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc.

Disruption and delay Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays.

Project A project is a temporary endeavor undertaken to create a unique product or service [1].

Definition of the Subject

There are many examples of complex projects suffering massive time and cost overruns. If a project has suffered such an overrun there may be a need to understand why it behaved the way it did. Two main reasons for this is (i) to gain learning for future projects or (ii) because one party of the project wishes to claim compensation from another party and thus is trying to explain what occurred during the project. In the latter case, system dynamics has been used for the last 30 years to help to understand why projects behave the way they do. Its success in this arena stems from its ability to model and unravel complex dynamic behavior that can result in project overruns. Starting from the first use of system dynamics in a claim situation in the late 1970's [2], it has directly influenced claim results worth millions of dollars. However, the number of claims which system dynamics has been involved in is still small as it is not perceived by project management practitioners as a standard tool for analyzing projects. System dynamics has a lot to offer in understanding complex projects, not only in a post-mortem situation, but it could also add value in the pre-project analysis stage and during the operational stage of a project.

Introduction

In this chapter we discuss the role of system dynamics (SD) modeling in understanding, and planning, a complex project. In particular we are interested in understanding how and why projects can go awry in a manner that seems surprising and often very difficult to unravel.

When we refer to projects we mean "a temporary endeavor undertaken to create a unique product or service" [1]. Projects are a specific undertaking, which implies that they are "one-shot", non-repetitive, time-limited, and, when complex, frequently bring about revolutionary (rather than evolutionary) improvements, start (to some extent) without precedent, and are risky with respect to customer, product, and project. If physical products are being created in a project, then the product is in some way significantly different to previous occasions of manufacturing (for example, in its engineering principles, or the expected operating conditions of the product, etc.), and it is this feature that means there is a need to take a project orientation.

Complex projects often suffer massive cost overruns. In recent decades those that have been publicized relate to large public construction projects, for example airports, bridges, and public buildings. Some examples include Denver's US$5 billion airport that was 200% overspent [3], the 800 million Danish Kroner Oresund bridge that was 68% overspent [4], and the UK's Scottish Parliament, which was 10 times the first budget [5]. The Major Projects Association [6] talks of a calamitous history of cost overruns of very large projects in the public sector. Flyvberg et al., [7] describe 258 major transportation infrastructure projects showing 90% of projects overspent.

Morris and Hough [8] conclude that "the track record of projects is fundamentally poor, particularly for the larger and more difficult ones. … Projects are often completed late or over budget, do not perform in the way expected, involve severe strain on participating institutions or are canceled prior to their completion after the expenditure of considerable sums of money." (p.7).

"Complex" projects are ones in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc. Complex projects, by their nature, comprise multiple interdependencies, and involve nonlinear relationships (which are themselves dynamic). For example, choices to accelerate might involve the use of additional overtime which can affect both learning curves and productivity as a result of fatigue – each of which are non-linear relationships. In addition many of the important features of complex projects are manifested through 'soft' relationships – for example managers will recognize deteriorating morale as projects become messy and look a failure, but assessing the impact of morale on levels of mistakes and rate of working has to be a matter of qualitative judgment. These characteristics are amenable particularly to SD modeling which specializes in working with qualitative relationships that are non-linear [9,10,11].

It is therefore surprising that simulation modeling has not been used more extensively to construct post-mortem analyzes of failed projects, and even more surprising because of SD's aptitude for dealing with feedback. Nevertheless the authors have been involved in the analysis of 10 projects that have incurred time and cost overruns and PA Consulting Group have claimed to have used SD to explain time and cost overruns for over 30 litigation cases [12]. Although in the mid-1990's, attempts to integrate SD modeling with more typical approaches to project management were emerging, their use has never become established within the project management literature or practice [13,14,15]. In addition, recognition that the trend towards tighter project delivery and accelerated development times meant that parallelism in project tasks was becoming endemic, and the impact of increasing parallelism could result in complex feedback dynamics where vicious cycles exist [16]. These vicious cycles are often the consequence of actions taken to enforce feedback control designed to bring a project back on track.

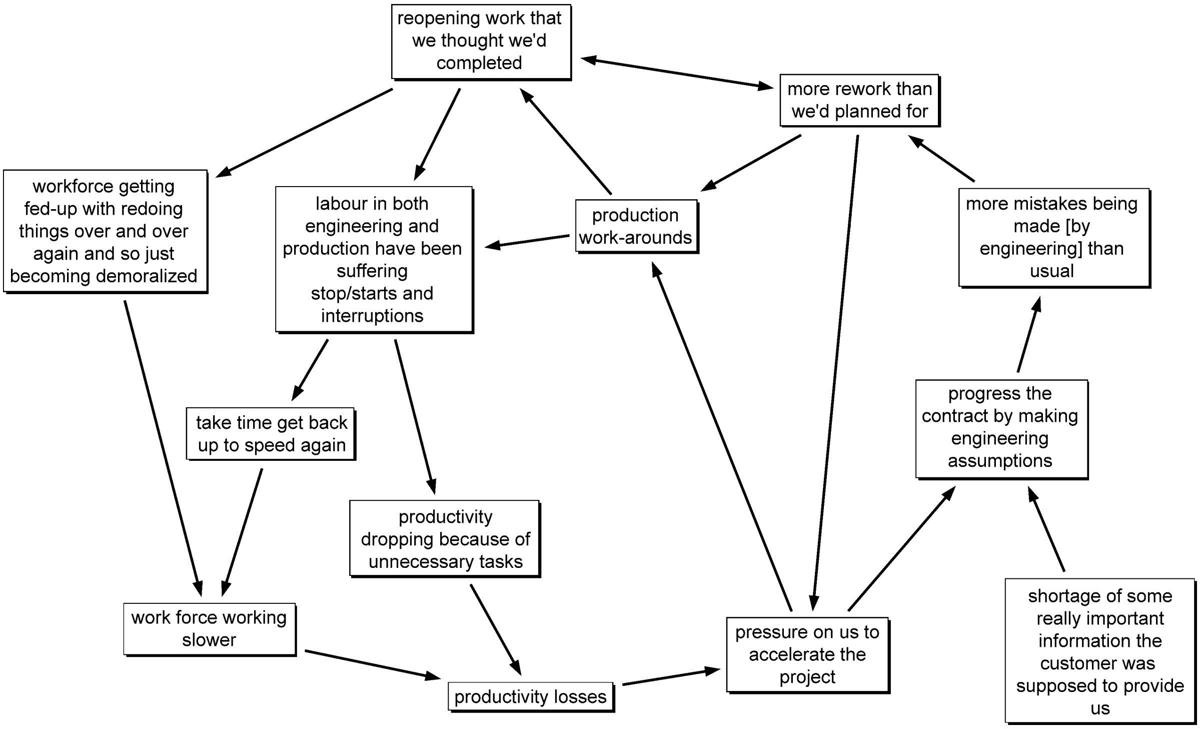
As project managers describe their experiences of projects going wrong they will often talk of these "vicious cycles" occurring, particularly with respect to the way in which customer changes seem to generate much more rework than might be expected, and that the rework itself then generates even more rework. Consider a small part of a manager's description of what he sees going on around him:

> "For some time now we've been short of some really important information the customer was supposed to provide us. As a consequence we've been forced to progress the contract by making engineering assumptions, which, I fear, have led to more mistakes being made than usual. This started giving us more rework than we'd planned for. But, of course, rework on some parts of the project has meant reopening work that we thought we'd completed, and that, in turn has reopened even more past work. Engineering rework has led to the need for production work-arounds and so our labour in both engineering and production have been suffering stop/starts and interruptions – and each time this happens they take time to get back up to speed again. This has led to productivity dropping because of unnecessary tasks, let alone productivity losses from the workforce getting fed-up with redoing things over and over again and so just becoming demoralized and so working slower. Inevitably all the rework and consequential productivity losses have put pressure on us to accelerate the project forcing us to have to make more engineering assumptions and do work-arounds."

Figure 1 shows a 'cause map' of the arguments presented by this project manager – the words used in the map are those used by the project manager and the arrows represent the causality described by the manager. This description is full of vicious cycles (indeed there are 35 vicious cycles discussed – see Fig. 1) all triggered by a shortage of customer furnished information and resulting in the rework cycle [17,18,19,20,21] and the need to accelerate in order to keep the project on schedule. Using traditional project management models such as Critical Path Method/Network Analysis cannot capture any of the dynamics depicted in Fig. 1, but SD simulation modeling is absolutely appropriate [22].

So, why has SD modeling been so little used? Partly it is because in taking apart a failed project the purpose is usually associated with a contractor wishing to make a claim for cost-overruns. In these circumstances the traditions of successful claims and typical attitudes of courts tend to determine the approach used. A 'measured-mile' approach is common, where numerical simplicity replaces the need for a proper understanding [23].

It was not until the early 1980's that the use of simulation modeling became apparent from publications in the public-domain. The settlement of a shipbuilding claim


reopening work that we thought we'd completed
more rework than we'd planned for
workforce getting fed-up with redoing things over and over again and so just becoming demoralized
labour in both engineering and production have been suffering stop/starts and interruptions
production work-arounds
more mistakes being made [by engineering] than usual
take time get back up to speed again
progress the contract by making engineering assumptions
productivity dropping because of unnecessary tasks
work force working slower
pressure on us to accelerate the project
shortage of some really important information the customer was supposed to provide us
productivity losses

Delay and Disruption in Complex Projects, Figure 1
Cause map showing the interactions described by a project manager and illustrating the feedback loops resulting from the complex dynamics behavior of a project under duress. The *arrows* represent causality

[2] prompted interest in SD modeling and [24], in the same year, reported on the use of management science modeling for the same purpose. It was not surprising that this modeling for litigation generated interest in modeling where the purpose was oriented to learning about failed projects (indeed the learning can follow from litigation modeling [25], although it rarely does).

As Fig. 1 demonstrates, it is not easy to understand fully the complex dynamic behavior of a project under duress. Few would realize that 35 feedback loops are encompassed in the description that led to Fig. 1. Indeed one of the significant features of complex projects is the likelihood of underestimating the complexity due to the dynamics generated by disruptions. [6] has reported on the more specific difficulty of understanding feedback behavior and research in the field of managerial judgment reinforces the difficulties of biases unduly influencing judgment [27].

In the work presented here we presume that there is a customer and a contractor, and there is a bidding process usually involving considerations of liquidated damages for delays and possibly strategic reputational consequences for late delivery. Thus, we expect the project to have a clear beginning and an end when the customer (internal or external) signs off a contract. Finally, we do not explore the whole project business life cycle, but that part where major cost overruns occur: thus, we start our consideration when a bid is to be prepared, consider development and manufacturing or construction, but stop when the product of the project is handed over to the customer.

Thus, in this chapter we shall be concerned specifically with the use of SD to model the consequences of disruptions and delays. Often these disruptions are small changes to the project, for example design changes [28]. The work discussed here is the consequence of 12 years of constructing detailed SD simulation models of failed complex projects. The first significant case was reported in Ackermann et al. [10] and Bennett et al. [29]. In each case the prompt for the work was the reasonable prospect of the contractor making a successful claim for damages. In all the cases the claim was settled out of court and the simulation model played a key role in settling the dispute.

The chapter will firstly consider why modeling disruption and delay (D&D) is so difficult. It will discuss what is meant by the term D&D and the typical consequences of D&D. This will be examined using examples from real projects that have suffered D&D. The contribution of SD modeling to the analysis of D&D and thus to the explanation of project behavior will then be discussed. A process of modeling which has been developed over the last 12 years and one that provides a means of modeling and explaining project behavior will be introduced. This process involves constructing both qualitative cause maps and quantitative system dynamics models. The chapter will conclude by considering potential future developments for the use of SD in modeling complex projects.

Disruption and Delay

(The following contains excerpts from Eden at al. [22] which provides a full discussion on the nature of D&D).

The idea that small disruptions can cause serious consequences to the life of a major project, resulting in massive time and cost overruns, is well established. The terms 'disruption and delay' or 'delay and disruption' are also often used to describe what has happened on such projects. However, although justifying the direct impact of disruptions and delays is relatively easy, there has been considerable difficulty in justifying and quantifying the claim for the indirect consequences. Our experience from working on a series of such claims is that some of the difficulty derives from ambiguity about the nature of disruption and delay (D&D). We now consider what we mean by D&D before moving onto considering the types of consequences that can result from the impact of D&D.

What is a Disruption?

Disruptions are events that prevent the contractor completing the work as planned. Many disruptions to complex projects are planned for at the bid stage because they may be expected to unfold during the project. For example, some level of rework is usually expected, even when everything goes well, because there will always be 'normal' errors and mistakes made by both the contractor and client. The disruption and delay that follows would typically be taken to be a part of a risk factor encompassed in the base estimate, although this can be significantly underestimated [30]. However, our experience suggests that *there are other types of disruptions that can be significant in their impact and are rarely thought about during original estimating*. When these types of disruptions do occur, their consequences can be underestimated as they are often seen by the contractor as aberrations with an expectation that their consequences can be controlled and managed. The linkage between risk assessment and the risks as potential triggers of D&D is often missed [31]. Interferences with the flow of work in the project is a common disruption. For example, when a larger number of design comments than expected are made by the client an increased number of drawings need rework. However it also needs to be recognized that these comments could have been made by the contractors own methods engineering staff. In either case, the additional work needed to respond to these comments, increases the contractor's workload and thus requires management to take mitigating actions if they still want to deliver on time. These mitigating actions are usually regarded as routine and capable of easily bringing the contract back to plan, even though they can have complex feedback ramifications.

Probably one of the most common disruptions to a project comes when a customer or contractor causes changes to the product (a Variation or Change Order). For example, the contractor may wish to alter the product after engineering work has commenced and so request a direct change. However, sometimes changes may be made unwittingly. For example, a significant part of cost overruns may arise where there have been what might be called 'give-aways'. These may occur because the contractor's engineers get excited about a unique and creative solution and rather than sticking to the original design, produce something better but with additional costs. Alternatively, when the contractor and customer have different interpretations of the contract requirements unanticipated changes can occur. For example, suppose the contract asks for a door to open and let out 50 passengers in 2 minutes, but the customer insists on this being assessed with respect to the unlikely event of dominantly large, slow passengers rather than the contractor's design assumptions of an average person. This is often known as 'preferential engineering'. In both instances there are contractor and/or customer requested changes that result in the final product being more extensive than originally intended.

The following example, taken from a real project and originally cited in Eden et al. [30], illustrates the impact of a client induced change to the product:

> **Project 1:** The contract for a 'state of the art' train had just been awarded. Using well-established design principles – adopted from similar train systems – the contractor believed that the project was on track. However within a few months problems were beginning to emerge. The client team was behaving very differently from previous experience and using the contract specification to demand performance levels beyond that envisioned by the estimating team. One example of these performance levels emerged during initial testing, 6 months into the contract, and related to water tightness. It was

> discovered that the passenger doors were not sufficiently watertight. Under extreme test conditions a small (tiny puddle) amount of water appeared. The customer demanded that there must be no ingress of water, despite acknowledging that passengers experiencing such weather would bring in more water on themselves than the leakage.
>
> The contractor argued that no train had ever met these demands, citing that most manufacturers and operators recognized that a small amount of water would always ingress, and that all operators accepted this. Nevertheless the customer interpreted the contract such that new methods and materials had to be considered for sealing the openings. The dialog became extremely combative and the contractor was forced to redesign. An option was presented to the customer for their approval, one that would have ramifications for the production process. The customer, after many tests and after the verdict of many external experts in the field, agreed to the solution after several weeks. Not only were many designs revisited and changed, with an impact on other designs, but also the delays in resolution impacted the schedule well beyond any direct consequences that could be tracked by the schedule system (www.Primavera.com) or costs forecasting system.

What is a Delay?

Delays are any events that will have an impact on the final date for completion of the project. Delays in projects come from a variety of sources. One common source is that of the client-induced delay. Where there are contractual obligations to comment upon documents, make approvals, supply information or supply equipment, and the client is late in these contractually-defined duties, then there may be a client-induced delay to the expected delivery date (although in many instances the delay is presumed to be absorbed by slack). But also a delay could be self-inflicted: if the sub-assembly designed and built did not work, a delay might be expected.

The different types of client-induced delays (approvals, information, etc.) have different effects and implications. Delays in client approval, in particular, are often ambiguous contractually. A time to respond to approvals may not have been properly set, or the expectations of what was required within a set time may be ambiguous (for example, in one project analyzed by the authors the clients had to respond within n weeks – but this simply meant that they sent back a drawing after n weeks with comments, then after the drawing was modified, they sent back the same drawing after a further m weeks with more comments). Furthermore, excessive comments, or delays in comments can cause chains of problems, impacting, for example, on the document approval process with sub-contractors, or causing over-load to the client's document approval process.

If a delay occurs in a project, it is generally considered relatively straightforward to cost. However, ramifications resulting from delays are often not trivial either to understand or to evaluate. Let us consider a delay only in terms of the CPM (Critical Path Method), the standard approach for considering the effects of delays on a project [32]. The consequences of the delay depend on whether the activities delayed are on the Critical Path. If they *are* on the Critical Path, or the delays are sufficient to cause the activities to become on the critical path, it is conceptually easy to compute the effect as an Extension Of Time (EOT) [33]. However, even in this case there are complicating issues. For example; what is the effect on other projects being undertaken by the contractor? When this is not the first delay, then to which schedule does the term "critical path" refer? To the original, planned programme, which has already been changed or disrupted, or to the "as built", actual schedule? Opinions differ here. It is interesting to note that, "the established procedure in the USA [of using as-built CPM schedules for claims] is almost unheard of in the UK" [33].

If the delay is *not* on the Critical Path then, still thinking in CPM terms, there are only indirect costs. For example, the activities on the Critical Path are likely to be resource dependent, and it is rarely easy to hire and fire at will – so if non-critical activities are delayed, the project may need to work on tasks in a non-optimal sequence to keep the workforce occupied; this will usually imply making guesses in engineering or production, requiring later re-work, less productive work, stop/starts, workforce overcrowding, and so on.

The following example, taken from a real project, illustrates the impact of a delay in client furnished information to the project:

> **Project 2:** A state of the art vessels project had been commissioned which demanded not only the contractor meeting a challenging design but additionally incorporating new sophisticated equipment. This equipment was being developed in another country by a third party. The client had originally guaranteed that the information on the equipment would be provided within the first few months of the contract – time enough for the information

to be integrated within the entire design. However time passed and no detailed specifications were provided by the third party – despite continual requests from the contractor to the client.

As the project had an aggressive time penalty the contractor was forced to make a number of assumptions in order to keep the design process going. Further difficulties emerged as information from the third party trickled in demanding changes from the emerging design. Finally manufacturing which had been geared up according to the schedule were forced to use whatever designs they could access in order to start building the vessel.

Portfolio Effect of many Disruptions

It is not just the extent of the disruption or delay but the number of them which may be of relevance. This is particularly the case when a large number of the disruptions and/or delays impact immediately upon one another thus causing a portfolio of changes. These portfolios of D&D impacts result in effects that would probably not occur if only one or two impacts had occurred. For example, the combination of a large number of impacts might result in overcrowding or having to work in poor weather conditions (see example below). In these instances it is possible to identify each individual item as a contributory cause of extra work and delay but not easy to identify the combined effect.

The following example, taken from another real project, illustrates the impact of a series of disruptions to the project:

Project 3: A large paper mill was to be extended and modernized. The extension was given extra urgency by new anti-pollution laws imposing a limit on emissions being enacted with a strict deadline.

Although the project had started well, costs seemed to be growing beyond anything that made sense given the apparent minor nature of the disruptions. Documents issued to the customer for ‘information only’ were changed late in the process. The customer insisted on benchmarking proven systems, involving visits to sites working with experimental installations or installations operating under different conditions in various different countries. In addition there were many changes of mind about where equipment should be positioned and how certain systems should work. Exacerbating these events was the circumstance of both the customer’s and contractor’s engineers being co-located, leading to ‘endless’ discussions and meetings slowing the rate of both design and (later) commissioning.

Relations with the customer, who was seen by the contractor to be continually interfering with progress of the project, were steadily deteriorating. In addition, and in order to keep the construction work going, drawings were released to the construction team before being fully agreed. This meant that construction was done in a piecemeal fashion, often inefficiently (for example, scaffolding would be put up for a job, then taken down so other work could proceed, then put up in the same place to do another task for which drawings subsequently had been produced). As the construction timescale got tighter and tighter, many more men were put on the site than was efficient (considerable overcrowding ensued) and so each task took longer than estimated.

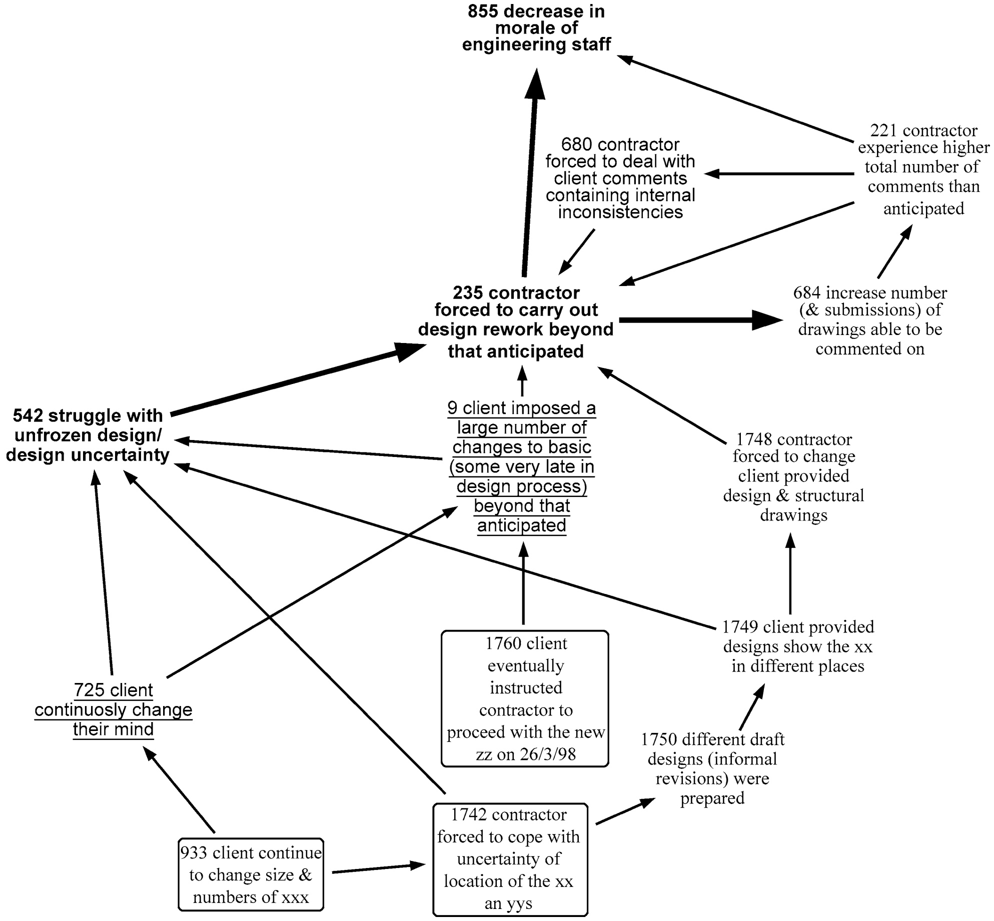
As a result the project was behind schedule, and, as it involved a considerable amount of external construction work, was vulnerable to being affected by the weather. In the original project plan (as used for the estimate) the outer shell (walls and roof) was due to be completed by mid-Autumn. However, the project manager now found himself undertaking the initial construction of the walls and roofing in the middle of winter! As chance would have it, the coldest winter for decades, which resulted in many days being lost while it was too cold to work. The combination of the particularly vicious winter and many interferences resulted in an unexpectedly huge increase in both labour hours and overall delay. Overtime payments (for design and construction workers) escalated. The final overspend was over 40% more than the original budget.

Consequences of Disruptions and Delays

Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays. If an unexpected variation (or disruption) occurs in a project then, if no intervention was to take place, a delivery delay would occur. In an attempt to avoid this situation, management may choose to take actions to prevent the delay (and possible penalties). In implementing these actions, side-effects can occur which cause further disruptions. These disruptions then cause further delays to the project. In order to avoid this situation, additional managerial action is required. Thus, an initial disruption has led to a delay, which has led to a disruption, which has led to a further delay. A positive feedback loop has been formed, where both disruption and delay feed back on themselves causing further disruptions and delays. Due to the nature of feedback loops, a powerful vicious cycle has been created which, if there is no alternative intervention, can escalate with the potential of getting 'out of control'. It is the dynamic behavior caused by these vicious cycles which can cause severe disruption and consequential delay in a project.

The dynamic behavior of the vicious cycles which are responsible for much of the D&D in a project make the costing of D&D very difficult. It is extremely difficult to separate each of the vicious cycles and evaluate their individual cost. Due to the dynamic behavior of the interactions between vicious cycles, the cost of two individual cycles will escalate when they interact with one another, thus disruptions have to be costed as part of a portfolio of disruptions.

Returning to Project 2, the vessel case, as can be seen in Fig. 2, the client caused both disruptions (continuous changes of mind) and delays (late permission to use a particular product). Both of these caused the contractor to undertake rework, and struggle with achieving a frozen (fixed) design. These consequences in turn impacted upon staff morale and also developed as noted above dynamic behavior – where rework resulted in more submissions of designs, which led to further comments, some of which were inconsistent and therefore led to further rework. As mentioned in the introduction, the rework


855 decrease in morale of engineering staff
680 contractor forced to deal with client comments containing internal inconsistencies
221 contractor experience higher total number of comments than anticipated
235 contractor forced to carry out design rework beyond that anticipated
684 increase number (& submissions) of drawings able to be commented on
542 struggle with unfrozen design/ design uncertainty
9 client imposed a large number of changes to basic (some very late in design process) beyond that anticipated
1748 contractor forced to change client provided design & structural drawings
1749 client provided designs show the xx in different places
1760 client eventually instructed contractor to proceed with the new zz on 26/3/98
725 client continuosly change their mind
1750 different draft designs (informal revisions) were prepared
1742 contractor forced to cope with uncertainty of location of the xx an yys
933 client continue to change size & numbers of xxx

Delay and Disruption in Complex Projects, Figure 2
Excerpt from a cause map showing some of the consequences of disruption and delay in Project 2. *Boxed* statements are specific illustrations with statements *underlined* representing generic categories (e. g. changes of mind). Statements in *bold text* represent the SD variables with the remainder providing additional context. *All links* are causal however those in *bold illustrate sections* of a feedback loop. *The numbers* at the beginning of concept are used as reference numbers in the model

cycle [17,18,19,20,21] can be a major driver of escalating feedback within a complex project.

Managerial Actions and the Consequences of D&D

The acceleration of disrupted projects to avoid overall project delays is common practice by managers who are under pressure from the client and/or their own senior management to deliver on time. However, the belief that this action will always help avoid delays is naive as it does not take into account an appreciation of the future consequences that can be faced. For example, one typical action designed to accelerate a project is to hire new staff. In doing so, some of the difficulties which may follow are:

- New staff take time to become acquainted with both the project and thus their productivity is lower than that of an existing skilled worker.
- New staff require training on the project and this will have an impact on the productivity of existing staff.
- Rather than hiring new staff to the organization, staff may be moved from other parts of the organization. This action results in costs to other projects as the other project is short of staff and so may have to hire workers from elsewhere, thereby suffering many of the problems discussed above.

Many of the outcomes of this action and other similar actions can lead to a reduction in expected productivity levels. Low productivity is a further disruption to the project through a lack of expected progress. If management identifies this lack of progress, then further managerial actions may be taken in an attempt to avoid a further delay in delivery. These actions often lead to more disruptions, reinforcing the feedback loop that had been set up by the first actions.

Two other common managerial actions taken to avoid the impact of a disruption on delivery are (i) the use of overtime and (ii) placing pressure on staff in an attempt to increase work rate. Both of these actions can also have detrimental effects on staff productivity once they have reached particular levels. Although these actions are used to increase productivity levels, effects on fatigue and morale can actually lead to a lowering of productivity via a slower rate of work and/or additional work to be completed due to increased levels of rework [21,34]. This lowering of productivity causes a delay through lack of expected progress on the project, causing a further delay to delivery. Management may then attempt to avoid this by taking other actions which in turn cause a disruption which again reinforces the feedback loop that has been set up.

Analyzing D&D and Project Behavior

The above discussion has shown that whilst D&D is a serious aspect of project management, it is a complicated phenomenon to understand. A single or a series of disruptions or delays can lead to significant impacts on a project which cannot be easily thought through due to human difficulties in identifying and thinking through feedback loops [26,35]. This makes the analysis of D&D and the resulting project behavior particularly difficult to explain.

SD modeling has made a significant contribution to increasing our understanding of why projects behave in the way they do and in quantifying effects. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project's behavior to the other (and, usually, why the actions of the other party has caused the project to behave in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project. In the case of a claim situation, although it has been shown that SD modeling can meet criteria for admissibility to court [36], there are a number of objectives which SD, or any modeling method, would need to address [37]. These include the following:

1. Prove causality – show what events triggered the D&D and how the triggers of D&D caused time and cost overruns on the project.
2. Prove the 'quantum' – show that the events that caused D&D created a specific time and cost over-run in the project. Therefore, there is a need to replicate over time the hours of work due to D&D that were over-and-above those that were contracted, but were required to carry out the project.
3. Prove responsibility – show that the defendant was responsible for the outcomes of the project. Also to demonstrate the extent to which plaintiff's management of the project was reasonable and the extent that overruns could not have been reasonably avoided.
4. All of the above have to be proved in a way which will be convincing to the several stakeholders in a litigation audience.

Over the last 12 years the authors have developed a model building process that aims to meet each of these purposes. This process involves constructing qualitative models to aid the process of building the 'case' and thus help to prove causality and responsibility (purposes 1 and 3). In addition, quantitative system dynamics models are involved in order to help to prove the quantum (purpose 2). However, most importantly, the process provides a structured, transparent, formalized process from "real world" interviews to resulting output which enables multiple audiences, including multiple non-experts as well as scientific/expert audiences to appreciate the validity of the models and thus gain confidence in these models and the consulting process in which they are embedded (purpose 4). The process is called the 'Cascade Model Building Process'. The next section describes the different stages of the model building process and some of the advantages of using the process.

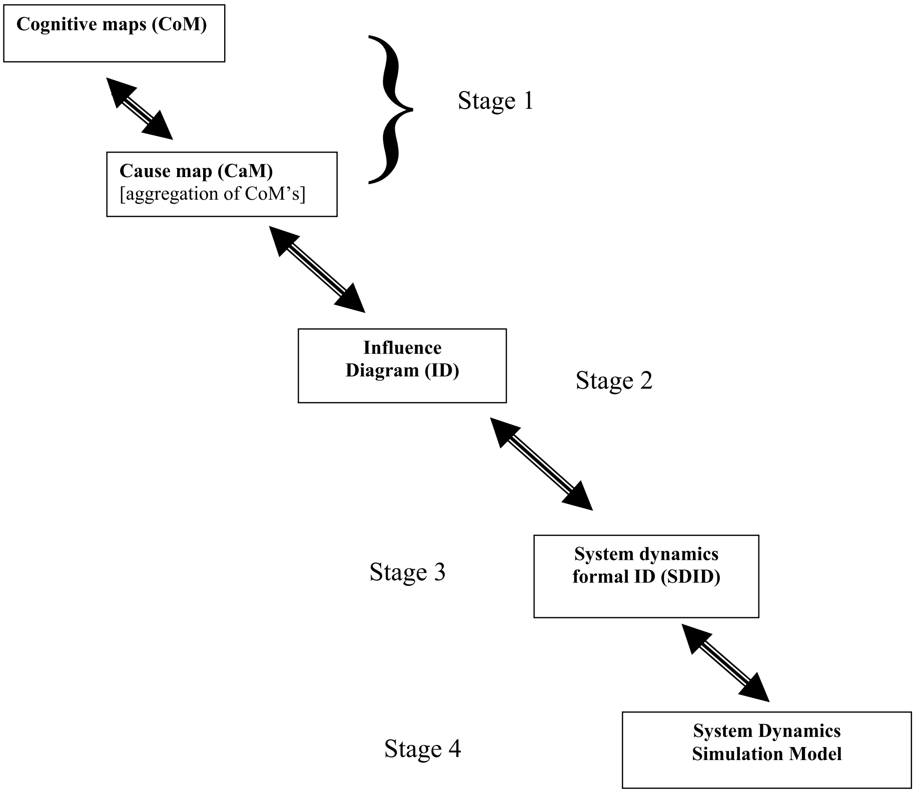
Cascade Model Building Process

(The following contains excerpts from Howick et al. [38], which contains a full description of the Cascade Model Building process).

The 'Cascade Model Building Process' involves four stages (see Fig. 3) each of which are described below.

Stage 1: Qualitative Cognitive and Cause Map

The qualitative cognitive maps and /or project cause map aim to capture the key events that occurred on the project, for example a delay as noted above in the vessel example in Project 2. The process of initial elicitation of these events can be achieved in two ways. One option is to interview, and construct cognitive maps [39,40,41] for each participant's views. Here the aim is to gain a deep and rich understanding that taps the wealth of knowledge of each individual. These maps act as a preface to getting the group together to review and assess the total content represented as a merged cause map [42] in a workshop setting. The second option is to undertake group workshops where participants can contribute directly, anonymously and simultaneously, to the construction of a cause map. The participants are able to 'piggy back' off one another, triggering new memories, challenging views and developing together a comprehensive overview [43]. As contributions from one participant are captured and structured to form a causal chain, this process triggers thoughts from others and as a result a comprehensive view begins to unfold. In Project 1, this allowed the relevant design engineers (not just those whose responsibility was the water tight doors, but also those affected who were dealing with car-body structure, ventilation, etc.), methods personnel and construction managers to surface a comprehensive view of the different events and consequences that emerged.

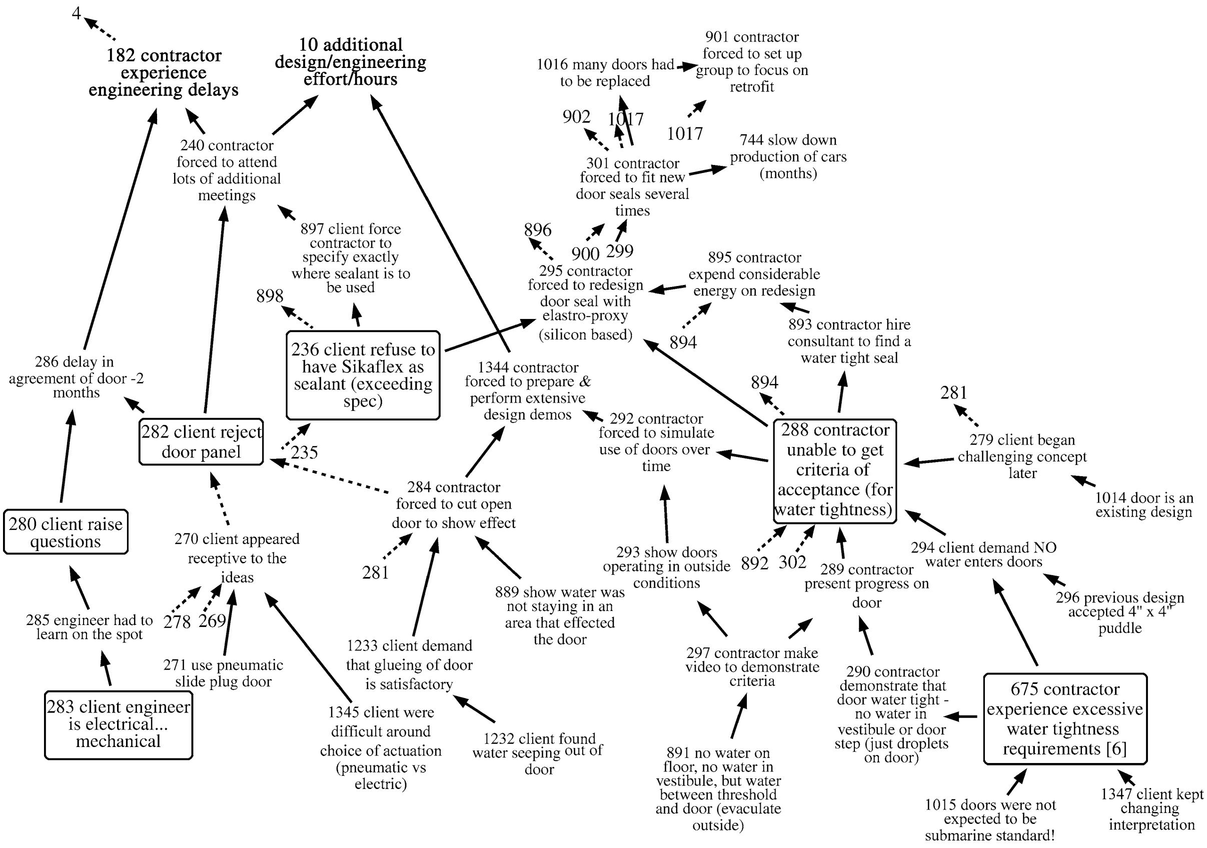


Delay and Disruption in Complex Projects, Figure 3
The Cascade Model Building Process

The continual development of the qualitative model, sometimes over a number of group workshops, engenders clarity of thought predominantly through its adherence to the coding formalisms used for cause mapping [44]. Members of the group are able to debate and consider the impact of contributions on one another. Through bringing the different views together it is also possible to check for coherency – do all the views fit together or are there inconsistencies? This is not uncommon as different parts of the organizations (including different discipline groups within a division e. g. engineering) encounter particular effects. For example, during an engineering project, manufacturing can often find themselves bewildered by engineering processes – why are designs so late. However, the first stage of the cascade process enables the views from engineering, methods, manufacturing, commissioning etc. to be integrated. Arguments are tightened as a result, inconsistencies identified and resolved and detailed audits (through analysis and features in the modeling software) undertaken to ensure consistency between both modeling team and model audience. In some instances the documents generated through reports about the organizational situation can be coded into a cause map and merged into the interview and workshop material [45].

The cause map developed at this stage is usually large – containing up to 1000 nodes. Computer supported analysis of the causal map can inform further discussion. For example, it can reveal those aspects of causality that are central to understanding what happened. Events that have multiple consequences for important outcomes can be detected. Feedback loops can be identified and examined. The use of software facilitates the identification of sometimes complex but important feedback loops that follow from the holistic view that arises from the merging of expertise and experience across many disciplines within the organization.

The resulting cause map from stage 1 can be of particular use in proving causality. For example, Fig. 4 repre-


4
182 contractor experience engineering delays
10 additional design/engineering effort/hours
1016 many doors had to be replaced
901 contractor forced to set up group to focus on retrofit
902
1017
1017
744 slow down production of cars (months)
240 contractor forced to attend lots of additional meetings
301 contractor forced to fit new door seals several times
897 client force contractor to specify exactly where sealant is to be used
896
900
299
295 contractor forced to redesign door seal with elastro-proxy (silicon based)
895 contractor expend considerable energy on redesign
898
894
893 contractor hire consultant to find a water tight seal
236 client refuse to have Sikaflex as sealant (exceeding spec)
286 delay in agreement of door -2 months
1344 contractor forced to prepare & perform extensive design demos
894
281
282 client reject door panel
235
292 contractor forced to simulate use of doors over time
288 contractor unable to get criteria of acceptance (for water tightness)
279 client began challenging concept later
1014 door is an existing design
284 contractor forced to cut open door to show effect
280 client raise questions
270 client appeared receptive to the ideas
281
293 show doors operating in outside conditions
892
302
289 contractor present progress on door
294 client demand NO water enters doors
296 previous design accepted 4" x 4" puddle
889 show water was not staying in an area that effected the door
285 engineer had to learn on the spot
278
269
1233 client demand that glueing of door is satisfactory
297 contractor make video to demonstrate criteria
290 contractor demonstrate that door water tight - no water in vestibule or door step (just droplets on door)
675 contractor experience excessive water tightness requirements [6]
271 use pneumatic slide plug door
283 client engineer is electrical... mechanical
1345 client were difficult around choice of actuation (pneumatic vs electric)
1232 client found water seeping out of door
891 no water on floor, no water in vestibule, but water between threshold and door (evaculate outside)
1015 doors were not expected to be submarine standard!
1347 client kept changing interpretation

Delay and Disruption in Complex Projects, Figure 4
Excerpt from a cause map showing some of the conversations regarding the water ingress situation in Project 1. As with Fig. 2, statements that have borders are the illustrations, those with *bold font* represent variables with the remainder detailing context. *Dotted arrows* denote the existence of further material which can be revealed at anytime

sents some of the conversations made regarding the water ingress situation described in the above case. In this figure, consequences such as additional engineering effort and engineering delays can be traced back to events such as client found water seeping out of door.

Stage 2: Cause Map to Influence Diagram

The causal model produced from stage 1 is typically very extensive. This extensiveness requires that a process of 'filtering' or 'reducing' the content be undertaken – leading to the development of an Influence Diagram (ID) (the second step of the cascade process). Partly this is due to the fact that many of the statements captured whilst enabling a detailed and thorough understanding of the project, are not relevant when building the SD model in stage 4 (as a result of the statements being of a commentary like nature rather than a discrete variable). Another reason is that for the most part SD models comprise fewer variables/auxiliaries to help manage the complexity (necessary for good modeling as well as comprehension).

The steps involved in moving from a cause map to an ID are as follows:

Step 1: Determining the core/endogenous variables of the ID

(i) Identification of feedback loops: It is not uncommon to find over 100 of these (many of these may contain a large percentage of common variables) when working on large projects with contributions from all phases of the project.

(ii) Analysis of feedback loops: Once the feedback loops have been detected they are scrutinized to determine a) whether there are nested feedback 'bundles' and b) whether they traverse more than one stage of the project. Nested feedback loops comprise a number of feedback loops around a particular topic where a large number of the variables/statements are common but with variations in the formulation of the feedback loop. Once detected, those statements that appear in the most number of the nested feedback loops are identified as they provide core variables in the ID model.

Where feedback loops straddle different stages of the process for example from engineering to manufacturing note is taken. Particularly interesting is where a feedback loop appears in one of the later stages of the project e. g. commissioning which links back to engineering. Here care must be taken to avoid chronological inconsistencies – it is easy to link extra engineering hours into the existing engineering variable however by the time commissioning discover problems in engineering, the majority if not all engineering effort has been completed.

Step 2: Identifying the triggers/exogenous variables for the ID The next stage of the analysis is to look for triggers – those statements that form the exogenous variables in the ID. Two forms of analysis provide clues which can subsequently be confirmed by the group:

(i) The first analysis focuses on starting at the end of the chains of argument (the tails) and laddering up (following the chain of argument) until a branch point appears (two or more consequences). Often statements at the bottom of a chain of argument are examples which when explored further lead to a particular behavior e. g. delay in information, which provides insights into the triggers.

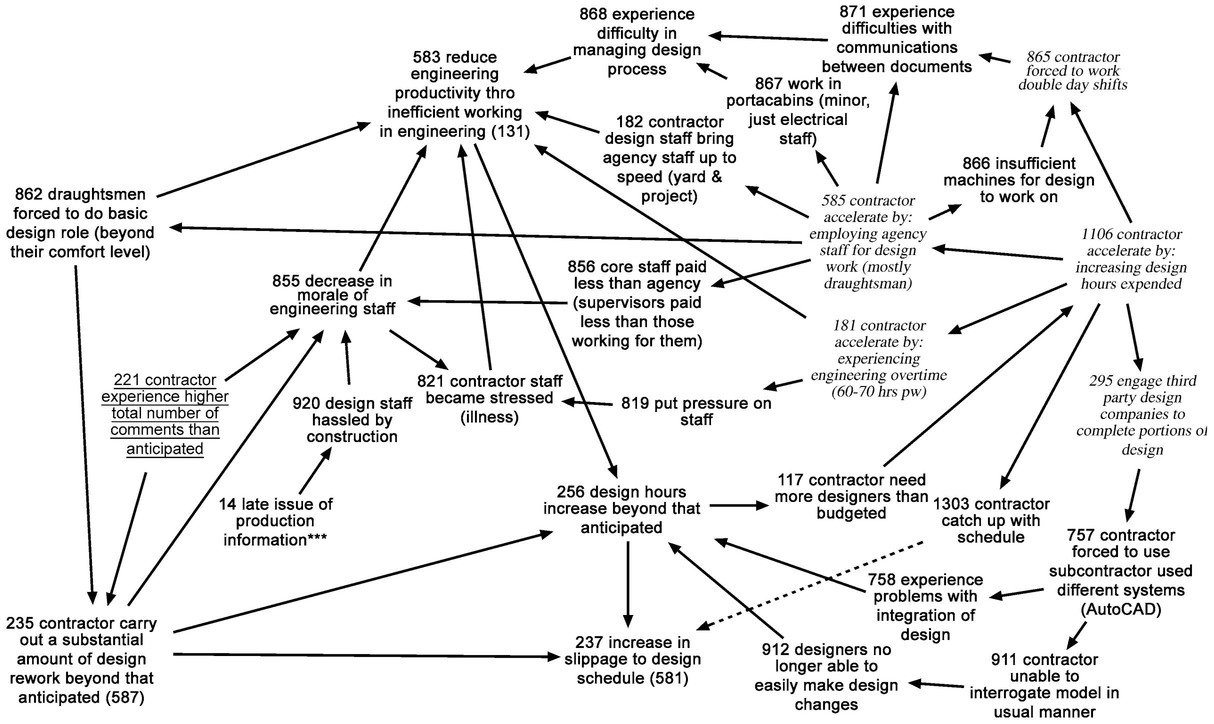
(ii) The initial set of triggers created by (i) can be confirmed through a second type of analysis – one which takes two different means of examining the model structure for those statements that are central or busy. Once these are identified they can be examined in more detail through creating hierarchical sets based upon them and thus "tear drops" of their content. Each of these teardrops is examined as possible triggers.

Step 3: Checking the ID Once the triggers and the feedback loops are identified care is taken to avoid double counting – where one trigger has multiple consequences some care must be exercised in case the multiple consequences are simple replications of one another.

The resulting ID is comparable to a 'causal loop diagram' [46] which is often used as a pre-cursor to a SD model. From the ID structure it is possible to create "stories" where a particular example triggers an endogenous variable which illustrates the dynamic behavior experienced.

Stage 3: Influence Diagram to System Dynamics Influence Diagram (SDID)

When a SD model is typically constructed after producing a qualitative model such as an ID (or causal loop diagram), the modeler determines which of the variables in the ID should form the stocks and flows in the SD model, then uses the rest of the ID to determine the main relationships that should be included in the SD model. However when building the SD model there will be additional


868 experience difficulty in managing design process
871 experience difficulties with communications between documents
865 contractor forced to work double day shifts
583 reduce engineering productivity thro inefficient working in engineering (131)
867 work in portacabins (minor, just electrical staff)
182 contractor design staff bring agency staff up to speed (yard & project)
866 insufficient machines for design to work on
862 draughtsmen forced to do basic design role (beyond their comfort level)
585 contractor accelerate by: employing agency staff for design work (mostly draughtsman)
1106 contractor accelerate by: increasing design hours expended
855 decrease in morale of engineering staff
856 core staff paid less than agency (supervisors paid less than those working for them)
181 contractor accelerate by: experiencing engineering overtime (60-70 hrs pw)
221 contractor experience higher total number of comments than anticipated
920 design staff hassled by construction
821 contractor staff became stressed (illness)
819 put pressure on staff
295 engage third party design companies to complete portions oi design
14 late issue of production information***
256 design hours increase beyond that anticipated
117 contractor need more designers than budgeted
1303 contractor catch up with schedule
757 contractor forced to use subcontractor used different systems (AutoCAD)
758 experience problems with integration of design
235 contractor carry out a substantial amount of design rework beyond that anticipated (587)
237 increase in slippage to design schedule (581)
912 designers no longer able to easily make design changes
911 contractor unable to interrogate model in usual manner

Delay and Disruption in Complex Projects, Figure 5
A small section of an ID from Project 2 showing mitigating actions (*italics*), triggers (*underline*) and some of the feedback cycles

variables/constants that will need to be included in order to make it 'work' that were not required when capturing the main dynamic relationships in the ID. The SDID is an influence diagram that includes all stocks, flows and variables that will appear in the SD model and is, therefore a qualitative version of the SD model. It provides a clear link between the ID and the SD model.

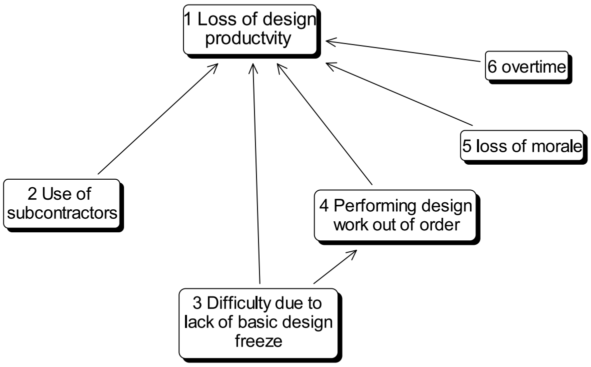
The SDID is therefore far more detailed than the ID and other qualitative models normally used as a pre-cursor to a SD model.

Methods have been proposed to automate the formulation of a SD model from a qualitative model such as a causal loop diagram [47,48,49] and for understanding the underlying structure of a SD model [50]. However, these methods do not allow for the degree of transparency required to enable the range of audiences involved in a claim situation, or indeed as part of an organizational learning experience, to follow the transition from one model to the next. The SDID provides an intermediary step between an ID and a SD model to enhance the transparency of the transition from one model to another for the audiences. This supports an auditable trail from one model to the next.

The approach used to construct the SDID is as follows: The SDID is initially created in parallel with the SD model. As a modeler considers how to translate an ID into a SD model, the SDID provides an intermediary step. For each variable in the ID, the modeler can do either of the following:

(i) Create one variable in the SD & SDID: If the modeler wishes to include the variable as one variable in the SD model, then the variable is simply recorded in both the SDID and the SD model as it appears in the ID.
(ii) Create multiple variables in the SD & SDID: To enable proper quantification of the variable, additional variables need to be created in the SD model. These variables are then recorded in both the SD model and SDID with appropriate links in the SDID which reflect the structure created in the SD model.

The SDID model forces all qualitative ideas to be placed in a format ready for quantification. However, if the ideas are not amenable to quantification or contradict one another, then this step is not possible. As a result of this process, a number of issues typically emerge including the need to add links and statements and the ability to assess the overall profile of the model though examining the impact of particular categories on the overall model structure. This process can also translate back into the causal model or ID model to reflect the increased understanding.


1 Loss of design productvity
6 overtime
5 loss of morale
2 Use of subcontractors
4 Performing design work out of order
3 Difficulty due to lack of basic design freeze

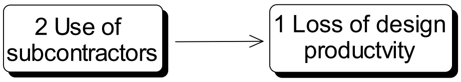
Delay and Disruption in Complex Projects, Figure 6
Section of an ID from Project 1 showing the factors affecting productivity

Stage 4: The System Dynamics Simulation Model

The process of quantifying SD model variables can be a challenge, particularly as it is difficult to justify subjective estimates of higher-level concepts such as "productivity" [51]. However, moving up the cascade reveals the causal structure behind such concepts and allows quantification at a level that is appropriate to the data-collection opportunities available. Figure 6, taken from the ID for Project 1, provides an example. The quantitative model will require a variable "productivity" or "morale", and the analyst will require estimation of the relationship between it and its exogenous and (particularly) endogenous causal factors. But while the higher-level concept is essential to the quantitative model, simply presenting it to the project team for estimation would not facilitate justifiable estimates of these relationships.

Reversing the Cascade

The approach of moving from stage 1 through to stage 4 can increase understanding and stimulate learning for all parties. However, the process of moving back up the cascade can also facilitate understanding between the parties. For example, in Fig. 7 the idea that a company was forced to use subcontractors and thus lost productivity might be a key part of a case for lawyers. The lawyers and the project team might have come at Fig. 7 as part of their construction of the case. Moving back up from the ID to the Cause Map (i. e. Fig. 7 to Fig. 8) as part of a facilitated discussion not only helps the parties to come to an agreed definition of the (often quite ill-defined) terms involved, it also helps the lawyers understand how the project team arrived at the estimate of the degree of the relationship. Having established the relationship, moving through the SDID (ensuring well-defined variables etc.) to the SD model enables the analysts to test the relationships to see whether any contradictions arise, or if model behaviors are significantly different from actuality, and it enables comparison of the variables with data that might be collected by (say) cost accountants. Where there are differences or contradictions, the ID can be re-inspected and if necessary the team presented with the effect of the relationship within the SD model explained using the ID, so that the ID and the supporting cause maps can be re-examined to identify the flaws or gaps in the reasoning. Thus, in this example, as simulation modelers, cost accountants, lawyers and engineers approach the different levels of abstraction, the cascade process provides a unifying structure within which they can communicate, understand each other, and equate terms in each others discourse.



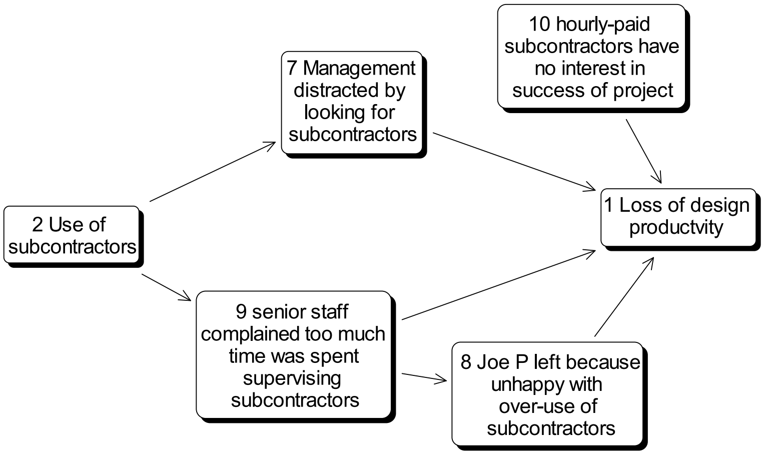
Delay and Disruption in Complex Projects, Figure 7
Section of an ID from Project 1 indicating the influence of the use of subcontractors on productivity

Advantages of the Cascade

The Cascade integrates a well-established method, cause mapping, with SD. This integration results in a number of important advantages for modeling to explain project behavior:

Achieving Comprehensiveness Our experience suggests that one of the principal benefits of using the cascade process derives from the added value gained through developing a rich and elaborated qualitative model that provides the structure (in a formalized manner) for the quantitative modeling. The cascade process immerses users in the richness and subtlety that surrounds their view of the projects and ensures involvement and ownership of all of the qualitative and quantitative models. The comprehensiveness leads to a better understanding of what occurred, which is important due to the complex nature of D&D, and enables effective conversations to take place across different organizational disciplines.

The process triggers new contributions as memories are stimulated and both new material and new connections are revealed. The resultant models thus act as organizational memories providing useful insights into future project management (both in relation to bids and implementation). These models provide more richness and therefore an increased organizational memory when compared to the traditional methods used in group model


10 hourly-paid subcontractors have no interest in success of project
7 Management distracted by looking for subcontractors
1 Loss of design productvity
2 Use of subcontractors
9 senior staff complained too much time was spent supervising subcontractors
8 Joe P left because unhappy with over-use of subcontractors

Delay and Disruption in Complex Projects, Figure 8
Section of a Cause Map from Project 1 explaining the relationship between the use of subcontractors and productivity

building for system dynamics models (for example [52]). However this outcome is not untypical of other problem structuring methods [53].

Testing the Veracity of Multiple Perspectives The cascade's bi-directionality enabled the project team's understandings to be tested both numerically and from the perspective of the coherency of the systemic portrayal of logic. By populating the initial quantitative model with data [10] rigorous checks of the validity of assertions were possible.

In a claim situation, blame can be the fear of those participating in accounting for history and often restricts contributions [44]. When initiating the cascade process, the use of either interviews or group workshops increases the probability that the modeling team will uncover the rich story rather than partial explanations or as is often the case with highly politicized situations, 'sanitized' explanations. By starting with 'concrete' events that can be verified, and exploring their multiple consequences, the resultant model provides the means to reveal and explore the different experiences of various stakeholders in the project.

Modeling Transparency By concentrating the qualitative modeling efforts on the capture and structuring of multiple experiences and viewpoints the cascade process initially uses natural language and rich description as the medium which facilitates generation of views and enables a more transparent record to be attained.

There are often insightful moments as participants viewing the *whole* picture realize that the project is more complex than they thought. This realization results in two advantages. The first is a sense of relief that they did not act incompetently given the circumstances i. e. the consequences of D&D took over – which in turn instills an atmosphere more conducive to openness and comprehensiveness (see [44]). The second is learning – understanding the whole, the myriad and interacting consequences and in particular the dynamic effects that occurred on the project (that often acts in a counter-intuitive manner) provides lessons for future projects.

Common Understanding Across many Audiences Claim situations involve numerous stakeholders, with varying backgrounds. The cascade process promotes ownership of the models from this mixed audience. For example, lawyers are more convinced by the detailed qualitative argument presented in the cause map (stage 1) and find this part of greatest utility and hence engage with this element of the cascade. However, engineers get more involved in the construction of the quantitative model and evaluating the data encompassed within it.

A large, detailed system dynamics model can be extremely difficult to understand for many of the stakeholders in a claim process [54]. However, the rich qualitative maps developed as part of the cascade method are presented in terms which are easier for people with no modeling experience to understand. In addition, by moving back up the cascade, the dynamic results that are output by the simulation model are given a grounding in the key events of the project, enabling the audience to be given fuller explanations and reasons for the D&D that occurred on the project.

Using the cascade method, any structure or parameters that are contained in the simulation model can be easily, and quickly, traced back to information gathered as a part of creating the cognitive maps or cause maps. Each contribution in these maps can then normally be traced to an individual witness who could defend that detail in the model. This auditable trail can aid the process of explaining the model and refuting any attacks made on the model.

Clarity The step-by-step process forces the modeler to be clear in what statements mean. Any illogical or inconsistent statements highlighted, require the previous stage to be revisited and meanings clarified, or inconsistencies cleared up. This results in clear, logical models.

Confidence Building As a part of gaining overall confidence in a model, any audience for the model will wish to have confidence in the structure of the model (for example [55,56,57,58]). When assessing confidence levels in a part of the structure of a SD model, the cascade process enables any member of the 'client' audience to clearly trace the structure of the SD model directly to the initial natural language views and beliefs provided from individual interviews or group sessions.

Scenarios are also an important test in which the confidence of the project team in the model can be considerably strengthened. Simulation is subject to the demands to reproduce scenarios that are recognizable to the managers capturing a portfolio of meaningful circumstances that occur at the same time, including many qualitative aspects such as morale levels. For example, if a particular time-point during the quantitative simulation is selected, clearly the simulated values of all the variables, and in particular the relative contributions of factors in each relationship, can be output from the model. If we consider Fig. 6, the simulation might show that at a particular point in a project, loss of productivity is 26% and that the loss due to:

"Use of subcontractors" is 5%.

"Difficulty due to lack of basic design freeze" is 9%.

"Performing design work out of order" is 3%.

"loss of morale" is 5%.

"overtime" is 4%.

Asking the project team their estimates of loss of productivity at this point in time, and their estimation of the relative contribution of these five factors, will help to validate the model. In most cases this loss level is best captured by plotting the relative levels of productivity against the time of critical incidents during the life the project. Discussion around this estimation might reveal unease with the simple model described in Fig. 6, which will enable discussion around the ID and the underlying cause map, either to validate the agreed model, or possibly to modify it and return up the cascade to further refine the model. In this scenario, validation of the cascade process provides a unifying structure within which the various audiences can communicate and understand each other.

The Cascade Model Building Process provides a rigorous approach to explaining why a project has behaved in a certain way. The cascade uses rich, qualitative stories to give a grounding in the key events that drive the behavior of the project. In addition, it provides a quantifiable structure that allows the over time dynamics of the project to be described. The Cascade has therefore contributed significantly in understanding why projects behave in the way they do.

This chapter has focused on the role of SD modeling in explaining the behavior of complex projects. The final two sections will consider the implications of this work and will explore potential future directions for the use of SD modeling of projects.

Implications for Development

So what is the current status of SD modeling of projects? What is the research agenda for studying projects using SD? Below we consider each aspect of the project life-cycle in turn, to suggest areas where SD modeling may be applied, and to consider where further work is needed.

The first area is pre-project risk analysis. Risk analysis traditionally looks at risks individually, but looking at the systemicity in risks has clear advantages [59]. Firstly, the use of cause mapping techniques by an experienced facilitator, aided by software tools, is a powerful means of drawing out knowledge of project risk from an individual manager (or group of managers), enhancing clarity of thought, allowing investigation of the interactions between risks, and enhancing creativity. It is particularly valuable when used with groups, bringing out interactions between the managers and helping to surface cultural differences. And it clearly enables analysis of the systemicity, in particular identification of feedback dynamics, which can help explicate project dynamics in the ways discussed above. The influence of such work has led to the ideas of cause maps, influence diagrams and SD to be included into risk practice standard advice (the UK "PRAM" Guide, edition 2 [60] – absent from Edition 1). In one key example [31], the work described above enabled the team to develop a 'Risk Filter' in a large multi-national project-based organization, for identifying areas of risk exposure on future projects and creating a framework for their investigation. The team reviewed the system after a few years; it had been used by 9 divisions, on over 60 major projects, and completed by 450 respondents; and it was used at several stages during the life of a project to aid in the risk assessment and contribute to a project database. The system allowed investigation of the interactions between risks, and so encouraged the management of the causality of relationships between risks, rather than just risks, thus focusing attention on those risks and causality that create the most frightening ramifications on clusters of risks, as a system, rather than single items. This also encouraged conversations about risk mitigation across disciplines within the organization. Clearly cause mapping is useful in risk analysis, but there are a number of research questions that follow, for example:

- In looking at possible risk scenarios, what are appropriate methodologies to organize and facilitate heterogeneous groups of managers? And how technically can knowledge of systemicity and scenarios be gathered into one integrated SD model and enhance understanding? [61]
- How can SD models of possible scenarios be populated to identify key risks? How does the modeling cascade help in forward-looking analysis?
- There are many attempts to use Monte-Carlo simulation to model projects, without taking the systemic issues into account – leading to models which can be seriously misleading [62]. SD models can give a much more realistic account of the effect of risks – but how can essentially deterministic SD models as described above be integrated into a stochastic framework to undertake probabilistic risk analyzes of projects which acknowledges the systemicity between the risks and the systemic effects of each risk?
- The use of SD is able to identify structures which give projects a propensity for the catastrophic systemic effects discussed in the Introduction. In particular, the three dimensions of structural complexity, uncertainty, and severe time-limitation in projects can combine together to cause significant positive feedback. However, defining metrics for such dimensions still remains an important open question. While a little work has been undertaken to give operational measures to the first of these (for example [63,64]), and de Meyer et al. [65] and Shenhar and Dvir [66] suggest selecting the management strategy based on such parameters, there has been little success so far in quantifying these attributes. The use of the SD models discussed above needs to be developed to a point where a project can be parametrized to give quantitatively its propensity for positive feedback.
- Finally, SD modeling shows that the effects of individual risks can be considerably greater than intuition would indicate, and the effects of clusters of risks particularly so. How can this be quantified so that risks or groups of risks can be ranked in importance to provide prioritization to managers? Again, Howick et al. [61] gives some initial indications here, but more work is needed.

The use of SD in operational control of projects has been less prevalent (Lyneis et al., [12] refers to and discusses examples of where it has been used). For a variety of reasons, SD and the traditional project management approach do not match well together. Traditional project-management tools look at the project in its decomposed pieces in a structured way (networks, work breakdown structures, etc.); they look at operational management problems at a detailed level; SD models aggregate into a higher strategic level and look at the underlying structure and feedback. Rodrigues and Williams [67] describe one attempt at an integrated methodology, but there is scope for research into how work with the SD paradigm can contribute to operational management of projects, and Williams [68] provides some suggestions for hybrid methods.

There is also a more fundamental reason why SD models do not fit in easily into conventional project management. Current project management practice and discourse is dominated by the "Bodies of Knowledge" or BoKs [69], which professional project management bodies consider to be the core knowledge of managing projects [1,70], presenting sets of normative procedures which appear to be self-evidently correct. However, there are three underlying assumptions to this discourse [71].

- Project Management is self-evidently correct: it is rationalist [72] and normative [73].
- The ontological stance is effectively positivist [74].
- Project management is particularly concerned with managing scope in individual parts [75].

These three assumptions lead to three particular *emphases* in current project management discourse and thus in the BoKs [71]:

- A heavy emphasis on planning [73,76].
- An implication of a very conventional control model [77].
- Project management is generally decoupled from the environment [78].

The SD modeling work provided explanations for why some projects severely over-run, which clash with these assumptions of the current dominant project management discourse.

- Unlike the third assumption, the SD models show behavior arising from the complex interactions of the various parts of the project, which would *not* be predicted from an analysis of the individual parts of the project [79].

- Against the first assumption, the SD models show project behavior which is complex and non-intuitive, with feedback exacerbated through management response to project perturbations, conventional methods provide unhelpful or even disbeneficial advice and are not necessarily self-evidently correct.
- The second assumption is also challenged. Firstly, the models differ from the BoKs in their emphasis on, or inclusion of, "soft" factors, often important links in the chains of causality. Secondly, they show that the models need to incorporate not only "real" data but management perceptions of data and to capture the socially constructed nature of "reality" in a project.

The SD models tell us why failures occur in projects which exhibit complexity [63] – that is, when they combine *structural complexity* [80] – many parts in complex combinations – and *uncertainty*, in project goals and in the means to achieve those goals [81]. Goal uncertainty in particular is lacking in the conventional project management discourse [74,82], and it is when uncertainty affects a structurally complex traditionally-managed project that the systemic effects discussed above start to occur. But there is a third factor identified in the SD modeling. Frequently, events arise that compromise the plan at a faster rate than that at which it is practical to re-plan. When the project is heavily *time-constrained*, the project manager feels forced to take acceleration actions. A structurally complex project when perturbed by external uncertainties can become unstable and difficult to manage, and under time-constraints dictating acceleration actions when management has to make very fast and sometimes very many decisions, the catastrophic over-runs described above can occur. Work from different direction seeking to establish characteristics that cause complexity projects come up with similar characteristics (for example [66]). But the SD modeling explains *how* the tightness of the time-constraints strengthen the power of the feedback loops which means that small problems or uncertainties can cause unexpectedly large effects; it also shows how the type of under-specification identified by Flyvberg et al. [4] brings what is sometimes called "double jeopardy" – under-estimation (when the estimate is elevated to the status of a project control-budget) which leads to acceleration actions that then cause feedback which causes much greater over-spend than the degree of under-estimation.

Because of this, the greatest contribution that SD has made – and perhaps can make – is to increase our understanding of why projects behave in the way they do. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project's behavior to the others (and, usually, why the actions of the other party has caused the project to behave in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project.

The bulk of the work referred to in this chapter comes in the first of these, the claim situation. However, while these have proved popular amongst SD modelers, they have not necessarily found universal acceptance amongst the practicing project-management community. Work is needed therefore in a number of directions. These will be discussed in the next section.

Future Directions

We have already discussed the difficulty that various audiences can have in comprehending a large, detailed system dynamics model [54], and that gradual explanations that can be given by working down (and back up) the cascade to bring understanding to a heterogeneous group (which might include jurors, lawyers, engineers and so on) and so link the SD model to key events in the project. While this is clearly effective, more work is needed to investigate the use of the cascade. In particular, ways in which the cascade can be most effective in promoting understanding, in formalizing the methodology and the various techniques mentioned above to make it replicable, as well as how best to use SD here (Howick [54] outlines nine particular challenges the SD modeler faces in such situations). Having said this, it is still the case that many forums in which claims are made are very set in conventional project-management thinking, and we need to investigate more how the SD methods can be combined with more traditional methods synergistically, so that each supports the other (see for example [83]).

Significant unrealized potential of these methodologies are to be found in the post-project "lessons learned" situation. Research has shown many problems in learning generic lessons that can be extrapolated to other projects, such as getting to the root causes of problems in projects, seeing the underlying systemicity, and understanding the narratives around project events (Williams [84], which gives an extensive bibliography in the area). Clearly, the modeling cascade, working from the messiness of individual perceptions of the situation to an SD model, can help in these areas. The first part of the process (Fig. 3), working through to the cause map, has been shown to enhance understanding in many cases; for example, Robertson and

Williams [85] describe a case in an insurance firm, and Williams [62] gives an example of a project in an electronics firm, where the methodology was used very "quick and dirty" but still gave increased understanding of why a (in that case successful) project turned out as it did, with some pointers to lessons learned about the process. However, as well as formalization of this part of the methodology and research into the most effective ways of bringing groups together to form cause maps, more clarity is required as to how far down the cascade to go and the additional benefits that the SD modeling brings. "Stage 4" describes the need to look at quantification at a level that is appropriate to the data-collection opportunities available, and there might perhaps be scope for SD models of parts of the process explaining particular aspects of the outcomes. Attempts to describe the behavior of the whole project at a detailed level may only be suitable for the claims situation; there needs to be research into what is needed in terms of Stages 3 and 4 for gaining lessons from projects (or, if these Stages are not carried out, how the benefits such as enhanced clarity and validity using the cause maps, can be gained).

One idea for learning lessons from projects used by the authors, following the idea of simulation "learning labs", was to incorporate learning from a number of projects undertaken by one particular large manufacturer into a simulation learning "game" [25]. Over a period of 7 years, several hundred Presidents, Vice-Presidents, Directors and Project Managers from around the company used the simulation tool as a part of a series of senior management seminars, where it promoted discussion around the experience and the effects encountered, and encouraged consideration of potential long-term consequences of decisions, enabling cause and effect relationships and feedback loops to be formed from participants' experiences. More research is required here as to how such learning can be made most effective.

SD modeling has brought a new view to project management, enabling understanding of the behavior of complex projects that was not accessible with other methods. The chapter has described methodology for where SD has been used in this domain. This last part of the chapter has looked forward to a research agenda into how the SD work needs to be developed to bring greater benefits within the project-management community.

Bibliography

1. Project Management Institute (2000) A guide to the Project Management Body of Knowledge (PMBOK). Project Management Institute, Newtown Square
2. Cooper KG (1980) Naval ship production: a claim settled and a framework built. Interfaces 10:20–36
3. Szyliowicz JS, Goetz AR (1995) Getting realistic about megaproject planning: the case of the new Denver International Airport. Policy Sci 28:347–367
4. Flyvberg B, Bruzelius N, Rothengatter W (2003) Megaprojects and risk: an anatomy of ambition. Cambridge University Press, Cambridge
5. Scottish Parliament (2003) Corporate body issues August update on Holyrood. Parliamentary News Release 049/2003
6. Major Projects Association (1994) Beyond 2000: A source book for major projects. Major Projects Association, Oxford
7. Flyvberg B, Holm MK, Buhl SL (2002) Understanding costs in public works projects: error or lie? J Am Plan Assoc 68:279–295
8. Morris PWG, Hough GH (1987) The anatomy of major projects. A study of the reality of project management. Wiley, Chichester
9. Forrester J (1961) Industrial dynamics. Productivity Press, Portland
10. Ackermann F, Eden C, Williams T (1997) Modeling for litigation: mixing qualitative and quantitative approaches. Interfaces 27:48–65
11. Lyneis JM, Ford DN (2007) System dynamics applied to project management: a survey, assessment, and directions for future research. Syst Dyn Rev 23:157–189
12. Lyneis JM, Cooper KG, Els SA (2001) Strategic management of complex projects: a case study using system dynamics. Syst Dyn Rev 17:237–260
13. Ford DN (1995) The dynamics of project management: an investigation of the impacts of project process and coordination on performance. Massachusetts Institute of Technology, Boston
14. Rodrigues A, Bowers J (1996) The role of system dynamics in project management. Int J Proj Manag 14:213–220
15. Rodrigues A, Bowers J (1996) System dynamics in project management: a comparative analysis with traditional methods. Syst Dyn Rev 12:121–139
16. Williams TM, Eden C, Ackermann F (1995) The vicious circles of parallelism. Int J Proj Manag 13:151–155
17. Cooper KG (1993) The rework cycle: benchmarks for the project manager. Proj Manag J 24:17–21
18. Cooper KG (1993) The rework cycle: How it really works.. and reworks… PMNETwork VII:25–28
19. Cooper KG (1993) The rework cycle: why projects are mismanaged. PMNETwork VII:5–7
20. Cooper KG (1993) The rework cycle: benchmarks for the project manager. Proj Manag J 24:17–21
21. Cooper KG (1994) The $2,000 hour: how managers influence project performance through the rework cycle. Proj Manag J 25:11–24
22. Eden C, Williams TM, Ackermann F, Howick S (2000) On the nature of disruption and delay. J Oper Res Soc 51:291–300
23. Eden C, Ackermann F, Williams T (2004) Analysing project cost overruns: comparing the measured mile analysis and system dynamics modelling. Int J Proj Manag 23:135–139
24. Nahmias S (1980) The use of management science to support a multimillion dollar precedent-setting government contact litigation. Interfaces 10:1–11
25. Williams TM, Ackermann F, Eden C, Howick S (2005) Learning from project failure. In: Love P, Irani Z, Fong P (eds) Knowledge management in project environments. Elsevier, Oxford

26. Sterman JD (1989) Modelling of managerial behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35:321–339
27. Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
28. Williams TM, Eden C, Ackermann F, Tait A (1995) The effects of design changes and delays on project costs. J Oper Research Society 46:809–818
29. Bennett PG, Ackermann F, Eden C, Williams TM (1997) Analysing litigation and negotiation: using a combined methodology. In: Mingers J, Gill A (eds) Multimethodology: the theory and practice of combining management science methodologies. Wiley, Chichester, pp 59–88
30. Eden C, Ackermann F, Williams T (2005) The amoebic growth of project costs. Proj Manag J 36(2):15–27
31. Ackermann F, Eden C, Williams T, Howick S (2007) Systemic risk assessment: a case study. J Oper Res Soc 58(1):39–51
32. Wickwire JM, Smith RF (1974) The use of critical path method techniques in contract claims. Public Contract Law J 7(1):1–45
33. Scott S (1993) Dealing with delay claims: a survey. Int J Proj Manag 11(3):143–153
34. Howick S, Eden C (2001) The impact of disruption and delay when compressing large projects: going for incentives? J Oper Res Soc 52:26–34
35. Diehl E, Sterman JD (1995) Effects of feedback complexity on dynamic decision making. Organ Behav Hum Decis Process 62(2):198–215
36. Stephens CA, Graham AK, Lyneis JM (2005) System dynamics modelling in the legal arena: meeting the challenges of expert witness admissibility. Syst Dyn Rev 21:95–122.35
37. Howick S (2003) Using system dynamics to analyse disruption and delay in complex projects for litigation: Can the modelling purposes be met? J Oper Res Soc 54(3):222–229
38. Howick S, Eden C, Ackermann F, Williams T (2007) Building confidence in models for multiple audiences: the modelling cascade. Eur J Oper Res 186:1068–1083
39. Eden C (1988) Cognitive mapping: a review. Eur J Oper Res 36:1–13
40. Ackermann F, Eden C (2004) Using causal mapping: individual and group: traditional and new. In: Pidd M (ed) Systems modelling: theory and practice. Wiley, Chichester, pp 127–145
41. Bryson JM, Ackermann F, Eden C, Finn C (2004) Visible thinking: unlocking causal mapping for practical business results. Wiley, Chichester
42. Shaw D, Ackermann F, Eden C (2003) Approaches to sharing knowledge in group problem structuring. J Oper Res Soc 54:936–948
43. Ackermann F, Eden C (2001) Contrasting single user and networked group decision support systems. Group Decis Negot 10(1):47–66
44. Ackermann F, Eden C, Brown I (2005) Using causal mapping with group support systems to elicit an understanding of failure in complex projects: some implications for organizational research. Group Decis Negot 14(5):355–376
45. Eden C, Ackermann F (2004) Cognitive mapping expert views for policy analysis in the public sector. Eur J Oper Res 152:615–630
46. Lane (2000) Diagramming conventions in system dynamics. J Oper Res Soc 51(2):241–245
47. Burns JR (1977) Converting signed digraphs to Forrester schematics and converting Forrester schematics to differential equations. IEEE Trans Syst Man Cybern SMC 7(10):695–707
48. Burns JR, Ulgen OM (1978) A sector approach to the formulation of system dynamics models. Int J Syst Sci 9(6):649–680
49. Burns JR, Ulgen OM, Beights HW (1979) An algorithm for converting signed digraphs to Forrester's schematics. IEEE Trans Syst Man Cybern SMC 9(3):115–124
50. Oliva R (2004) Model structure analysis through graph theory: partition heuristics and feedback structure decomposition. Syst Dyn Rev 20(4):313–336
51. Ford D, Sterman J (1998) Expert knowledge elicitation to improve formal and mental models. Syst Dyn Rev 14(4):309–340
52. Vennix J (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
53. Rosenhead J, Mingers J (2001) Rational analysis for a problematic world revisited. Wiley, Chichester
54. Howick S (2005) Using system dynamics models with litigation audiences. Eur J Oper Res 162(1):239–250
55. Ackoff RL, Sasieni MW (1968) Fundamentals of operations research. Wiley, New York
56. Rivett P (1972) Principles of model building. Wiley, London
57. Mitchell G (1993) The practice of operational research. Wiley, Chichester
58. Pidd M (2003) Tools for thinking: modelling in management science. Wiley, Chichester
59. Williams TM, Ackermann F, Eden C (1997) Project risk: systemicity, cause mapping and a scenario approach. In: Kahkonen K, Artto KA (eds) Managing risks in projects. E & FN Spon, London, pp 343–352
60. APM Publishing Ltd (2004) Project risk analysis and management guide. APM Publishing Ltd, High Wycombe, Bucks
61. Howick S, Ackermann F, Andersen D (2006) Linking event thinking with structural thinking: methods to improve client value in projects. Syst Dyn Rev 22(2):113–140
62. Williams TM (2004) Learning the hard lessons from projects – easily. Int J Proj Manag 22(4):273–279
63. Williams TM (1999) The need for new paradigms for complex projects. Int J Proj Manag 17:269–273
64. Shenhar AJ (2001) One size does not fit all projects: exploring classical contingency domains. Manag Sci 47:394–414
65. De Meyer A, Loch CH, Rich MT (2002) Managing project uncertainty: from variation to chaos. MIT Sloan Mgmt Rev 43(2):60–67
66. Shenhar AJ, Dvir D (2004) How project differ and what to do about it. In: Pinto J, Morris P (eds) Handbook of managing projects. Wiley, New York, pp 1265–1286
67. Rodrigues A, Williams TM (1997) Systems dynamics in software project management: towards the development of a formal integrated framework. Eur J Inf Syst 6:51–66
68. Williams TM (2002) Modelling complex projects. Wiley, Chichester
69. Dixon M (ed) (2000) The Association for Project Management (APM) Body of Knowledge (BoK), 4th edn. Association for Project Management, High Wycombe
70. Stevens M (2002) Project management pathways. Association for Project Management, High Wycombe
71. Williams TM (2005) Assessing and building on project management theory in the light of badly over-run projects. IEEE Trans Eng Manag 52(4):497–508
72. Lundin RA (1995) Editorial: temporary organizations and project management. Scand J Mgmt 11:315–317

73. Packendorff J (1995) Inquiring into the temporary organization: new directions for project management research. Scand J Mgmt 11:319–333
74. Linehan C, Kavanagh D (2004) From project ontologies to communities of virtue. Paper presented at the 2nd International Workshop, Making projects critical, University of Western England, 13–14th December 2004
75. Koskela L, Howell G (2002) The theory of project management: explanation to novel methods. In: Proceedings 10th Annual Conference on Lean Construction, IGLC-10, August 2002, Gramado, Brazil
76. Koskela L, Howell G (2002) The underlying theory of project management is obsolete. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 293–301
77. Hodgson DE (2004) Project work: the legacy of bureaucratic control in the post-bureaucratic organization. Organization 11:81–100
78. Malgrati A, Damiani M (2002) Rethinking the new project management framework: new epistemology, new insights. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 371–380
79. Lindkvist L, Soderlund J, Tell F (1998) Managing product development projects: on the significance of fountains and deadlines. Org Stud 19:931–951
80. Baccarini D (1996) The concept of project complexity – a review. Int J Proj Manag 14:201–204
81. Turner JR, Cochrane RA (1993) Goals-and-methods matrix: coping with projects with ill defined goals and/or methods of achieving them. Int J Proj Manag 11:93–102
82. Engwall M (2002) The futile dream of the perfect goal. In: Sahil-Andersson K, Soderholm A (eds) Beyond project management: new perspectives on the temporary-permanent dilemma. Libe Ekonomi, Copenhagen Business School Press, Malmo, pp 261–277
83. Williams TM (2003) Assessing extension of time delays on major projects. Int J Proj Manag 21(1):19–26
84. Williams TM (2007) Post-project reviews to gain effective lessons learned. Project Management Institute, Newtown Square
85. Robertson S, Williams T (2006) Understanding project failure: using cognitive mapping in an insurance project. Proj Manag J 37(4):55–71

Diffusion of Innovations, System Dynamics Analysis of the

PETER M. MILLING[1], FRANK H. MAIER[2]
[1] Industrieseminar der Universität Mannheim, Mannheim University, Mannheim, Germany
[2] International University in Germany, Bruchsal, Germany

Article Outline

Glossary

Adopters The cumulated number of persons who have bought a product over time.

Diffusion The spread of a new product, process or concept in the market. The process of bringing innovation into wide use.

Invention The process of bringing new technology into being.

Innovator A customer with general interest in innovations making his buying decision independent of others.

Innovation The process of bringing new technology into use.

Installed base Installed base is defined as the amount of users in a network system.

Imitator An imitator buy a new product because he observed or communicated with customers who have already bought the product. The buying decision of imitators is influenced by the adoption of other customers.

Network effects A product is characterized by a network effect, if the utility of that product is a function of the installed base. The utility increases with the installed base.

Definition of the Subject

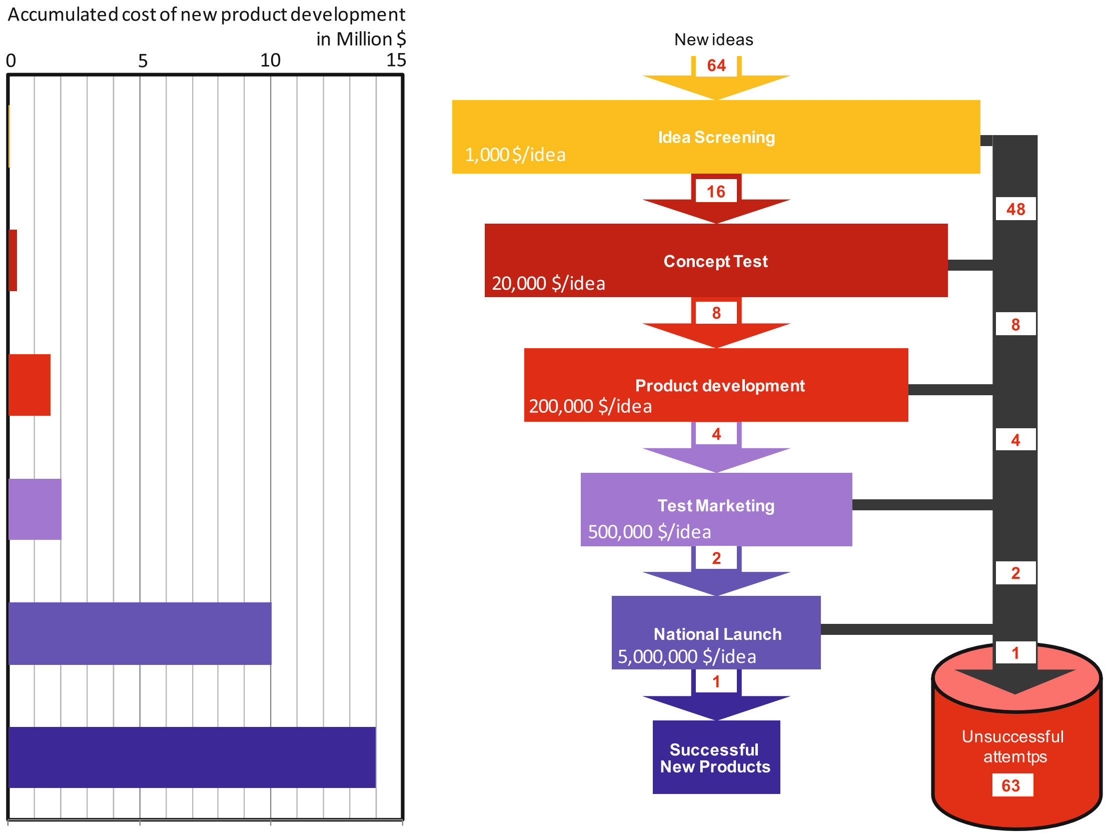
The article describes how system dynamics-based models can contribute to the understanding and improved management of the diffusion of innovations. It emphasizes the importance of an integrated feedback-oriented view of the different stages of innovation processes. The aim is to generate insight in the complexity and the dynamics of innovation processes. Based on the classical Bass model of innovation diffusion, the system dynamics perspective is introduced. In a systematic approach several structures to model the complexity and dynamics of managerial decision-making in the context of the diffusion of innovation are described and analyzed. Aspects covered consider market structure, network externalities, dynamic pricing, manufacturing related decisions and the link between research and development and the diffusion of a new product in the market place. The article concludes with managerial implications.

Introduction

Continuous activities to renew a company's range of products are crucial for the survival in a competitive environment. However, to improve the competitive position or the competitive advantage, ongoing innovation activity through the development, test, and introduction of new products is necessary. At least since the 1970s, it could be observed that new and technically more complex and sophisticated products have to be developed in a shorter time span. Resources have to be allocated to research and development (R&D) projects that are expected to be economically successful. New products have to be introduced to global markets with severe competition. Decisions about the adequate time to market and appropriate pricing, advertising, and quality strategies have to be made.

The complexity and difficulties to manage innovation activities partly derive from the comprehensiveness of the innovation processes. To be competitive, companies have to be successful in all stages of the innovation process, i. e., the process of invention, innovation, and diffusion. This becomes obvious when new product failure rates and innovation costs are analyzed. Figure 1 illustrates the cascading process of innovation activity and the related innovation costs.

For one successful new product in the market place, 64 promising ideas must be channeled through the process of invention and innovation. The cost at each stage of the invention and innovation process increases from a $1000 to $5 million per attempt. Not only is failure more expensive in later stages – which requires an effective management to reduce the failure rates – successful new products have to earn all necessary resources for the whole process. This requires the following: (1) to manage R&D projects and processes effectively and efficiently – including thorough and educated assessment of the economic potential



Diffusion of Innovations, System Dynamics Analysis of the, Figure 1
Outcome of activities along the process of invention and innovation

of a new product – to reduce failure rates in later stages and (2) to increase management attention in the final stages since failures in late stages of the process are much more expensive.

Models of innovation diffusion can support the complex and highly dynamic tasks. The article will briefly examine how system dynamics-based analysis of innovation diffusion can contribute to the understanding of the structures and forces driving the processes of innovation and diffusion. It will show how system dynamics models can support the decision-making and how they can help to reduce failures in the later stages of innovation activities.

Principle Structures to Model the Diffusion of Innovations

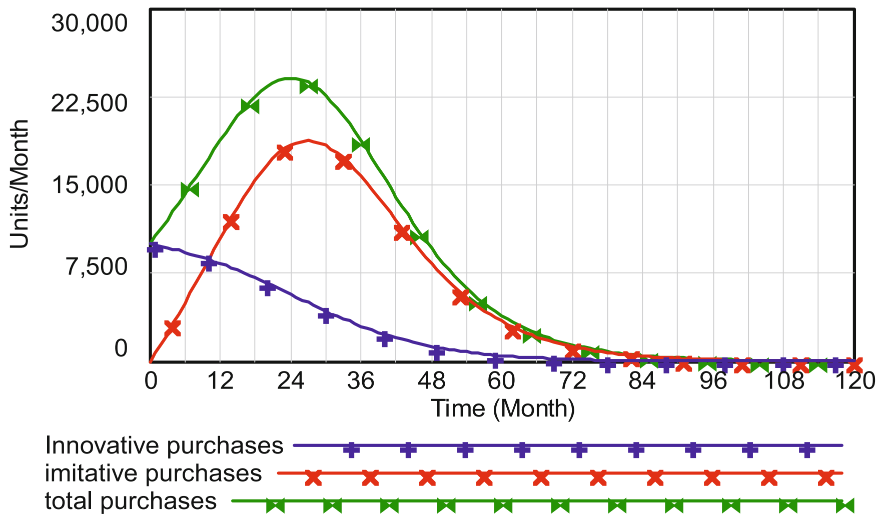
Traditional Innovation Diffusion Models from a System Dynamics Perspective

In literature discusses plenty of models about the diffusion of innovations. Many models are based on Frank M. Bass' model of innovation diffusion. In this model, product purchases result from two distinct forms of buying behavior, i. e., innovative purchases and imitative purchases. According to the original Bass model, innovative purchases of a period can be calculated as a fraction of the remaining market potential $(N - X_{t-1})$ with N being the market potential and $X_{t-1} = \sum_{\tau=0}^{t-1} S_\tau$ representing the accumulation of all past purchases of the product S_τ until period $t - 1$.

According to this, innovative purchases S_t^{inno} can be calculated as

$$S_t^{\text{inno}} = \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) \qquad (1)$$

where α represents the coefficient of innovation. In the original model, this coefficient is a constant essentially representing the fraction of innovators of the remaining market potential at any point of time. Imitative purchases, however, are influenced by the number of purchases in the past. Potential adopters of an innovation make their purchasing decision depending on the spread of the product



Diffusion of Innovations, System Dynamics Analysis of the, Figure 2
Product life cycle behavior generated by the Bass model

in the market place. The more customers have adopted the product in the past, the higher is the social pressure to purchase the product as well. Imitative demand of a period S_t^{imit} hence can be calculated as

$$S_t^{\text{imit}} = \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_\tau}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) \tag{2}$$

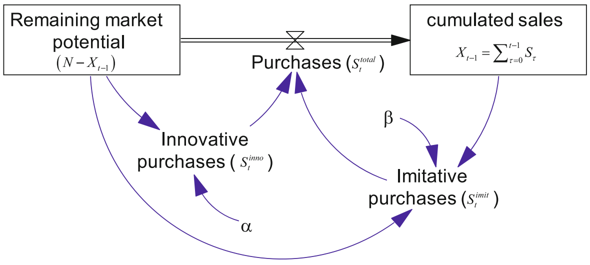
with β representing the coefficient of imitation – a probability that a purchase takes place by someone who observed the use of a product. Together, the total purchases in a period S_t^{total} equal $S_t^{\text{inno}} + S_t^{\text{imit}}$ and hence are calculated as

$$\begin{aligned} S_t^{\text{total}} = S_t^{\text{inno}} + S_t^{\text{imit}} &= \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) \\ &+ \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_\tau}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) . \end{aligned} \tag{3}$$

Innovative and imitative purchases together create the typical product life cycle behavior of the diffusion of an innovation in the market place as shown in Fig. 2.

The model above is a simple mathematical representation of the product life cycle concept, a key framework in business management. It describes the time pattern a product follows through subsequent stages of introduction, growth, maturity, and decline. Because of its mathematical simplicity and its ability to represent the diffusion of an innovation, the Bass model has been used for parameter estimation and therefore serves as a base for projections of future sales. Although the concept is a powerful heuristic, many models generating this typical behavior do not consider e. g., actual economic environment, competition, capital investment, cost and price effects. Innovation diffusion models, which do not comprise the relevant decision variables, exhibit a significant lack of policy content. They do not explain how structure conditions behavior. They cannot indicate how actions of a firm can promote but also impede innovation diffusion. For an improved understanding of innovation dynamics generated by feedback structures that include managerial decision variables or economic conditions, the system dynamics approach is highly suitable.

Equations (1) to (3) can easily be transformed into the system dynamics terminology. $(N - X_{t-1})$ represents the stock of the remaining market potential at any point in time and X_{t-1} represents the accumulation of all product purchases over time. The sales of a period S_t^{total} are the flows connecting these two stocks as shown in the Fig. 3.

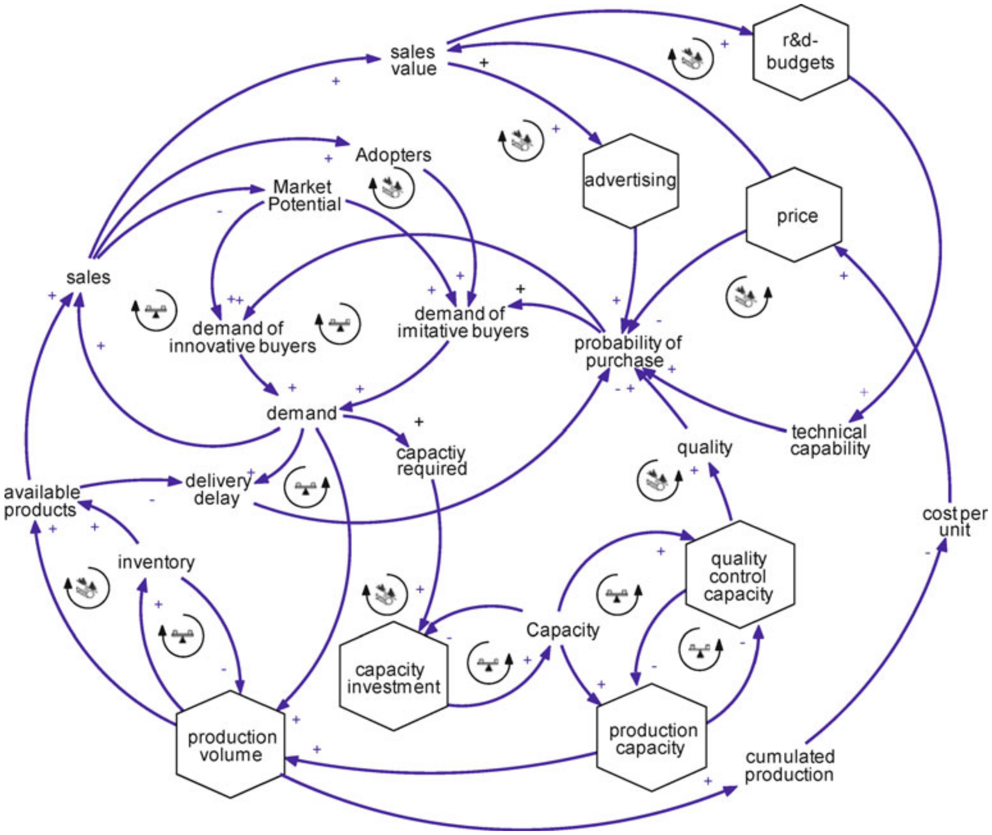

Remaining market potential
$(N - X_{t-1})$
Purchases (S_t^{total})
cumulated sales
$X_{t-1} = \sum_{\tau=0}^{t-1} S_\tau$
Innovative purchases (S_t^{inno})
β
Imitative purchases (S_t^{imit})
α

Diffusion of Innovations, System Dynamics Analysis of the, Figure 3
Stock-flow view of the Bass model

The coefficients α and β represent the probability of a purchase taking place; they are constants in the original Bass model and independent of any decisions or changes over time. For this reason, the model has been criticized and subsequent models have been developed that make the coefficients depending on variables like price or advertising budget. Most of the extensions, however, include no feedback between the diffusion process and these decision variables. This is a severe shortcoming since in the market place, diffusion processes are strongly influenced by feedback. What in classical innovation diffusions models typically is referred to as word-of-mouth processes is nothing else than a reinforcing feedback process. Adopters of an innovation – represented by the cumulated sales X_{t-1} – communicate with potential customers $(N - X_{t-1})$ and – by providing information about the product – influence their behavior. However, feedback in innovation diffusion goes beyond the pure word-of-mouth processes. It also involves the decision processes of a company and the outcome generated by the purchasing decision of the customers like the sales volume generated.

Figure 4 describes as a causal loop diagram the diversity of potential influences of corporate decision variables (marked with hexagons) on demand of the products by making the probability of a purchase – the coefficients α and β – depending on decision variables. It also shows how corporate decisions are interconnected through several feedback structures and influence the diffusion of a new product in the market place. Although being far from a comprehensive structure of potential feedback, the figure gives an impression of the complex dynamic nature of innovation diffusion processes.

Decision variables like pricing or advertising directly influence the purchase probability of a potential customer. The higher the advertising budgets and the lower the price, the higher will be demand for the products of a company. Furthermore, there are indirect and/or delayed effects on the speed of the spread of a new product in the market. Actual sales of a product may be limited by insuf-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 4
Feedback structures driving innovation processes

ficient production and inventory levels which increases delivery delays (perceived or actual) and therefore reduce demand. Growing demand, however, motivates the company to expand its capacity and to increase the volume of production. This leads to higher cumulated production and through experience curve effects to decreasing costs per unit, lower prices, and further increased demand. Other influences might reflect that a certain percentage of total available production capacity has to be allocated to ensure the quality of the output – either by final inspection or during the production process. Quality control then will improve product quality, which directly affects demand.

Models developed in this manner can serve as simulators to analyze the consequences of strategies and to improve understanding. They can show e. g., how pricing and investment strategies depend on each other and quantify the impact of intensified quality control on production and sales. They are suitable tools to investigate the effects resulting from the impact of a particular management problem on the dynamic complexity of innovation diffusion. Creating an understanding of the processes and interactions is the main purpose of system dynamics-based innovation diffusion models. Subsequently, a base structure of a system dynamics-based model will be described.

Base Structure of a System Dynamics-Based Model of Innovation Diffusion

First, a model will be discussed that maps the diffusion of an innovation in a monopolistic situation or can serve as an industry level model. Secondly, competition between potential and existing companies is introduced. Thirdly, substitution between successive product generations is considered. Each step adds complexity to the model. This approach allows for a better understanding of the forces driving the spread of a new product in the market.

In the following, the coarse structure of a model generating the life cycle in the market of a new product is presented and analyzed in its dynamic implications in Sect. "Representing Managerial Decision Making in Innovation Diffusion Models". Figure 5 gives an aggregated view the main model structure. It also introduces – in contrast to the mathematical terms known from the Bass model, variable names, which are informative and consistent with the use in system dynamics models.

The diffusion of a new product is generated by the behavior of the before mentioned two different types of buyers: innovators and imitators. If the potential customers (PC) – i. e., the remaining market potential of a product – decide to purchase, either as innovators or as imitators, they become adopters *(ADOP)*. The variables *PC* and *ADOP* and their associated transfer rates are the basic variables of the core diffusion process. The untapped market *(UM)* covers latent demand that can be activated by appropriate actions and leads to an increase in the number of potential customers and therefore increases the remaining market potential. Besides the growth resulting from the influx from the untapped market, a decline in market volume can be caused by the loss of potential customers to competitors. This lost demand *(LD)* turned to competing products that are more attractive, e. g., products of a higher level of technological sophistication, quality or lower price.

The differentiation into the two buying categories "innovators" and "imitators" refers to the Bass model of innovation diffusion as described in Subsect. "Traditional Innovation Diffusion Models from a system dynamics Perspective". The distinction is made because these two types of buyers react differently to prices charged, product quality offered, advertisements or the market penetration already achieved. The term "innovator" refers to customers who make their purchasing decision without being influenced by buyers who already purchased the product, the adopters. In the beginning of an innovation diffusion process, innovators take up the new product because they are interested in innovations. The number of innovators is a function of the potential customers. Mathematically, the purchasing decision of innovators D^{Inno} is defined by a coefficient of innovation α times the number of potential customers *PC*.

$$D^{\text{inno}}_{(t)} = \alpha_{(t)} \cdot PC_{(t)} \tag{4}$$

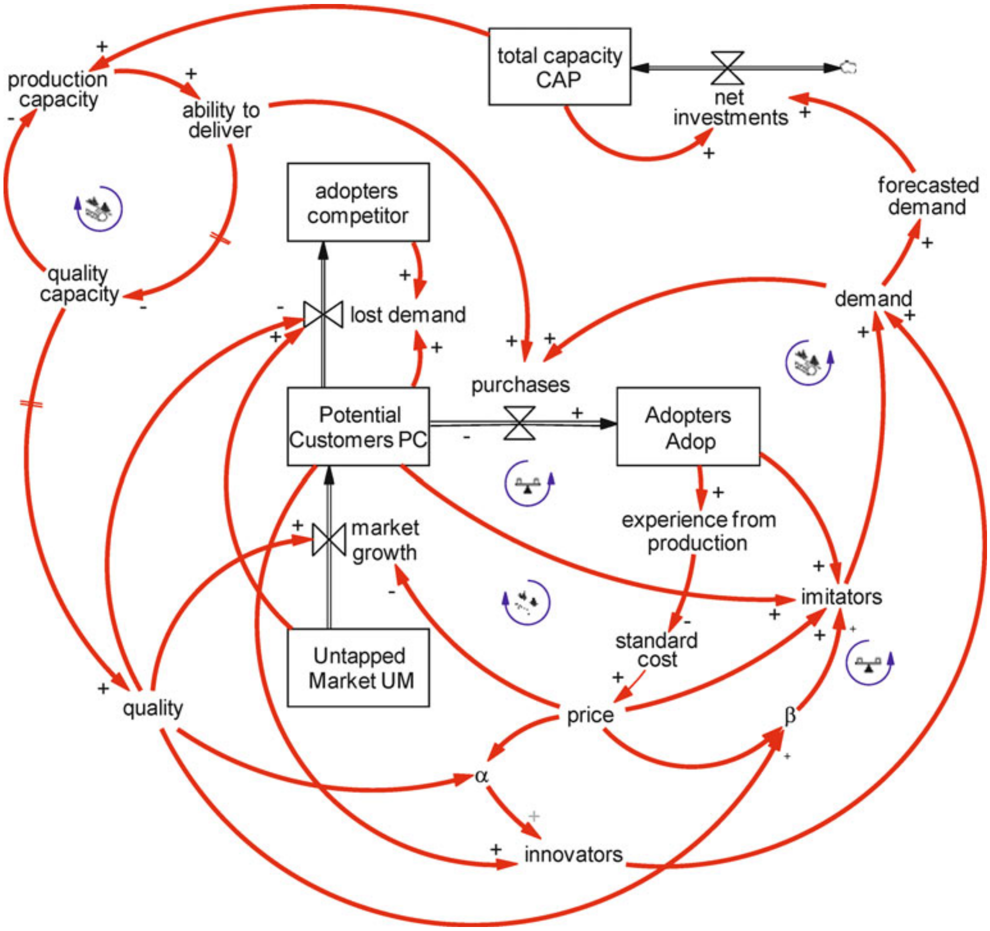
with:

$D^{\text{inno}}_{(t)}$ Demand from innovators

$\alpha_{(t)}$ Coefficient of innovation

$PC_{(t)}$ Potential customers .

The purchasing decision of "imitators" is calculated differently. Imitators buy a new product because they observe or communicate with customers who have already adopted the product. They imitate the observed buying behavior. Innovators initiate new product growth, but the diffusion gains momentum from the word-of-mouth process between potential customers and the increasing level of adopters. The driving force behind the imitation process is communication – either personal communication between an adopter and someone who still does not own the product or just observation of someone already owning and using the product. Although, the Bass model describes how the imitators' purchases can be calculated – as shown in Eq. (2) – the equation can also be derived from



Diffusion of Innovations, System Dynamics Analysis of the, Figure 5
Coarse structure of the innovation diffusion model

a combinatorial analysis of the number of possible contacts between the adopters and the potential customers. If N is the total number of people in a population consisting of potential customers PC and adopters $ADOP$, the amount of possible combinations C_N^k is

$$C_N^k = \binom{N}{k} = \frac{N!}{k!(N-k)!}\,. \tag{5}$$

Here we are only interested in paired combinations ($k = 2$) between the elements in N

$$\begin{aligned} C_N^2 = \binom{N}{2} &= \frac{N!}{2!(N-2)!} \\ &= \frac{N(N-1)}{2!} = \frac{1}{2}(N^2 - N)\,. \end{aligned} \tag{6}$$

Since N represents the sum of elements in PC and in $ADOP$, $(N = PC + ADOP)$, the number of combinations between potential customers and adopters is

$$\begin{aligned} &= \frac{1}{2}\left[(PC + ADOP)^2 - (PC + ADOP)\right] \\ &= \frac{1}{2}\left[PC^2 + 2 \cdot PC \cdot ADOP + ADOP^2 - PC - ADOP\right] \end{aligned} \tag{7}$$

and after regrouping and collecting terms, we get

$$\begin{aligned} = \frac{1}{2}\big(&\underbrace{2 \cdot PC \cdot ADOP}_{\substack{\text{Communication between} \\ PC \text{ and } ADOP}} + \underbrace{PC^2 - PC}_{\substack{\text{Communication} \\ \text{within } PC}} \\ &+ \underbrace{ADOP^2 - ADOP}_{\substack{\text{Communication} \\ \text{within } ADOP}}\big)\,. \end{aligned} \tag{8}$$

Internal communications, both within PC and $ADOP$, generate no incentive to purchase the new product and are neglected; the process of creating imitative buying decisions in Eq. (9) is, therefore, reduced to the first term in Eq. (8), the information exchange between potential customers and adopters.

$$D_{(t)}^{\text{imit}} = \beta^* \cdot PC_{(t)} \cdot ADOP_{(t)} \tag{9}$$

with:

$D_{(t)}^{\text{imit}}$ Demand from imitators

$\beta_{(t)}^*$ Coefficient of imitation $= \dfrac{\beta_{(t)}}{N}$

$ADOP_{(t)}$ Adopters

N Initial market potential.

The coefficient of imitation $\beta_{(t)}^*$ represents the original coefficient of innovation β from the Bass model divided by the initial market potential N. β can be interpreted as the probability that the possible contacts between members in *PC* and *ADOP* have been established, relevant information has been exchanged, and a purchasing decision is made.

The sum of the demand of innovators and imitators in each period, $D_{(t)}$, establishes the basic equation for the spread of a new product in the market. Together with the state variables of potential customers and adopters the flows of buyers (innovators and imitators) constitute the core model of innovation diffusion, which generates the typical s-shaped pattern of an adoption process over time.

$$\begin{aligned} D_{(t)} &= D_{(t)}^{\text{inno}} + D_{(t)}^{\text{imit}} \\ &= \alpha_{(t)} \cdot PC_{(t)} + \beta_{(t)}^* \cdot PC_{(t)} \cdot ADOP_{(t)} \,. \end{aligned} \tag{10}$$

Although Eqs. (3) and (10) are based on different interpretations and explanations, they are structurally identical since *PC* equals $(N - X_{t-1})$ and *ADOP* equals $X_{t-1} = \sum_{\tau=0}^{t-1} S_\tau$. The only difference is that the coefficients of innovation and imitation, in the context of the model based on (10) are now a variable – rather than a constant – depending on corporate decision variables like price or quality. Furthermore, corporate decisions are not just set as predefined time paths; they are endogenously calculated and depend on the outcome of the diffusions process itself. Model simulations of this extended innovation diffusion model will be discussed in Sect. "Representing Managerial Decision Making in Innovation Diffusion Models".

Extending the Base Structure to Include Competition

In the model described above, competition is not modeled explicitly. The model only assumes a potential loss in demand, if price, quality or ability to deliver are not within the customers' expectations. The internal corporate structures of competition are not explicitly represented. To generate diffusion patterns that are influenced by corporate decisions and the resulting dynamic interactions of the different competitors in a market, a more sophisticated way to incorporate competition is needed. Therefore, a subscript i $(i = 1, 2, \ldots, k)$ representing a particular company is introduced as a convenient and efficient way to model the different competitors. In a competitive innovation diffusion model the calculation of innovative and imitative demand of a company has to be modified. Equation (4) that determines the innovative demand in a monopolistic situation becomes Eq. (11) – in the following discussion, the time subscript *(t)* is omitted for simplicity. The coefficient of innovation α has to be divided by the number of competitors N to ensure that each company will have the same share of innovative demand as long as there is no differentiation among the competitors' products through, e. g., through pricing or advertising. The subscript i in the coefficient of innovation is necessary because it considers that the decisions of an individual company regarding product differentiation influences its proportion of innovative buyers. A third modification is necessary, because the number of competitors may vary over time. Therefore, the term φ_i represents a factor to model different dates of market entry. It takes the value 1 if a company i is present at the market, otherwise it is 0. Hence, the demand of company i is 0, as long as it is not present at the market and $\sum_{i=1}^{k} \varphi_i$ represents the actual number of competitors. The variable potential customers *PC* has no subscript because all companies in the market compete for a common group of potential customers, whereas innovative demand has to be calculated for each company.

$$D_i^{\text{inno}} = \frac{\alpha_i}{NC} \cdot PC \cdot \varphi_i \tag{11}$$

with:

α_i coefficient of innovation for company i

NC number of active competitors $= \displaystyle\sum_{i=1}^{k} \varphi_i$

φ_i factor of market presence company i

i subscript representing the companies $i = (1, 2, \ldots, k)$.

The buying decisions of imitators are influenced by observation of, or communication with the adopters *(ADOP)*. In a competitive environment two alternative approaches can be used to calculate imitative demand. These different approaches are a result of different interpretations of the object of the communication processes. In the first interpretation, the 'product related communication', the adopters of a particular company's product communicate information about the product they have purchased e. g., an electronic device like a MP3 player of a particular company. In this case, the calculation of imitative demand has to consider the number of potential contacts between the potential customers *PC* and the adopters of the products of company i ($ADOP_i$) as shown in Eq. (12).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot ADOP_i \cdot PC \cdot \varphi_i \qquad (12)$$

with:

β_i coefficient of imitation for company i.

The second interpretation about the object of communication is the 'product form-related communication'. Here, the adopters communicate information about a product form, for example, DVD players in general and not about an MP3 player of a particular company. The equation to calculate imitative demand for the model of product form related communication is shown in Eq. (13). The sum of adopters for each company i $\left(\sum_{i=1}^{k} ADOP_i\right)$ represents the total number of adopters in the market. The product of the total adopters and the potential customers then represents the total number of potential contacts in the market. Imitative demand of a company i depends on the share of total adopters $\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}$ this company holds.

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \frac{ADOP_i}{\sum_{i=1}^{k} ADOP} \cdot PC \cdot \sum_{i=1}^{k} ADOP_i \cdot \phi_i \,. \qquad (13)$$

If the term that represents a company's share of the total adopters of a market $\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}$ is raised to the power of γ as in Eq. (14), weaker ($0 < \gamma < 1$) or stronger ($\gamma > 1$) influences of a company's share of total adopters on demand can be represented explicitly. For $\gamma = 1$, Eq. (14) is identical to Eq. (13).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \left(\frac{ADOP_i}{\sum_{i=1}^{k} ADOP}\right)^{\gamma} \cdot PC \cdot \sum_{i=1}^{k} ADOP_i \cdot \phi_i \qquad (14)$$

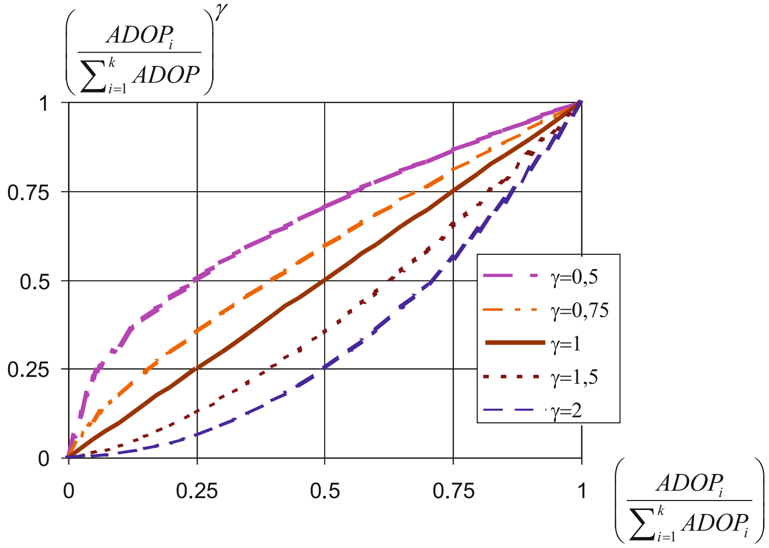
with:

γ factor representing customers' resistance to "Me-too"-pressure.

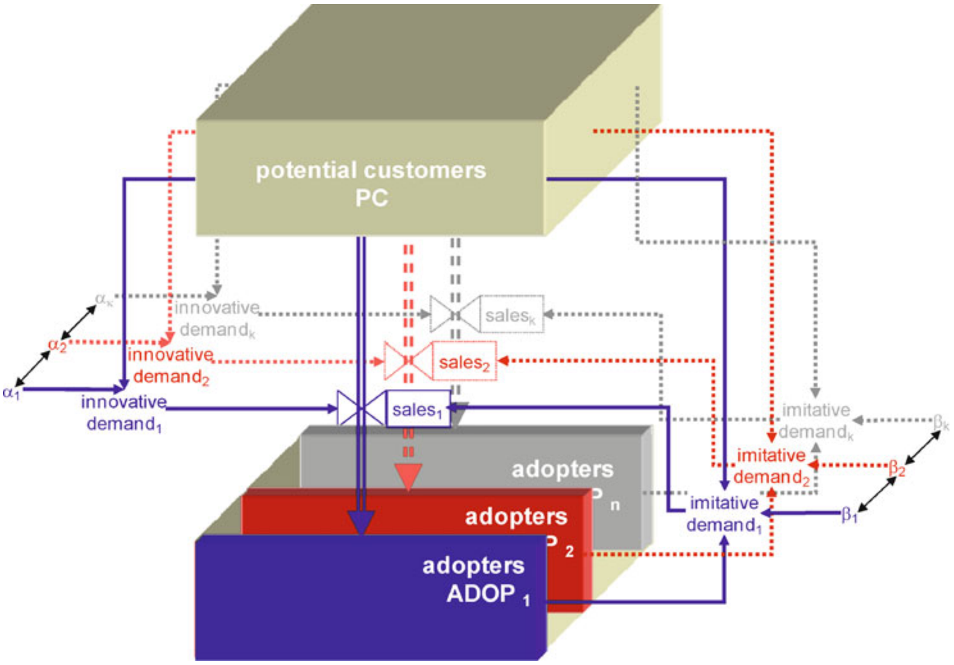
Figure 6 shows the effects of a company's share of the total adopters for different γ. For a given share of total adopters this means: the higher γ, the lower is the value of the term

$$\left(\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}\right)^{\gamma}$$

and the stronger is the importance of a high share of total adopters. The parameter γ can be interpreted as a measure



Diffusion of Innovations, System Dynamics Analysis of the, Figure 6
Effects of a company's share of adopters for different γ



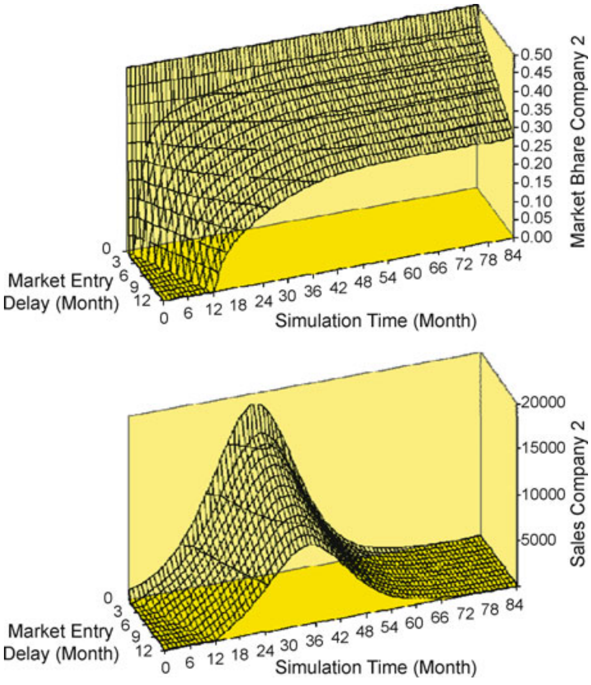
Diffusion of Innovations, System Dynamics Analysis of the, Figure 7
Coarse structure of an oligopolistic innovation diffusion model

of the importance of customer loyalty or as resistance to "me-too" pressure.

Figure 7 illustrates the coarse structure of an oligopolistic innovation diffusion model as described by Eqs. (11) and (14). The hexahedron at the top represents the stock of potential customers *PC* for the whole market. The blocks with the different shading represent for each company *i* the level of adopters, i. e., the cumulated sales of the company. The total number of adopters of the product form corresponds to the addition of these blocks.

Since the sales are calculated separately for each company *i* there are n outflows from the stock of potential customers to the adopters. Again, sales comprise innovative and imitative demand, which are influenced by the coefficient of innovation α_i and imitation β_i. Both coefficients are influenced by managerial decisions of each company *i* like pricing, advertising, quality, market entry timing, etc. and measure the relative influence of the decisions compared to the competitor's decisions. Therefore, the values α_i and β_i not only depend on the decisions of company *i*, they also depend on the competitor's decisions. Both variables are crucial for the speed and the maximum volume of demand for the products of a company *i*.

Figure 8 shows the results of simulations based on Eq. (11) for innovative demand and Eq. (14) for imitative demand with the effects of a market entry delay of



Diffusion of Innovations, System Dynamics Analysis of the, Figure 8
Follower's market share and sales for different market entry times

the second company – the influences of other decision variables are switched off. Several model simulations have been made assuming a market entry delay of company 2 between 0 and 12 months.

The plots in Fig. 8 show the development of market share and sales of the second company over time. Since there is no further product differentiation, both competitors have the same market share when they enter the market at the same time. With each month delay of the second company the market share that can be achieved at the end of the simulation decreases. A three months delay reduces the finally achieved market share to 40%; a 12-month delay even causes a decrease in market share down to approximately 25%. Accordingly, the maximum sales volume decreases significantly with each month delay in market entry time.

Representing Network Externalities

In the following, we will investigate the diffusion of a specific type of goods in order to show the importance of understanding the diffusion of goods with network effects (based on [22]). The trend towards an information society has stressed the relevance of goods satisfying information and communication needs. Many products of this market segment such as electronic mail contain attributes that necessitate a specific examination, since the diffusion of goods showing network effects differs from that of conventional products. The main difference between conventional products and products with network effects is that the utility of the latter cannot be regarded as a constant value. With regard to these products, utility is an endogenous variable which results in a specific diffusion behavior. Two effects are responsible for this particular behavior: the bandwagon effect and the penguin effect. A refined system dynamics model supports a better understanding of this special diffusion process.

The fact that the utility is not constant can be reasoned by a concept commonly referred to as "network effect". A product is characterized by a network effect, if the utility of that product is a function of the installed base, which is defined as the amount of users in a network system. The utility increases with the installed base. This leads to a virtual interdependency of the users inside the network. The interdependency is based on the bandwagon effect and the penguin effect. Starting from the fundamental diffusion model of Bass the characteristics of network effects are integrated into the model in order to simulate the diffusion behavior.

Many definitions for network effects (sometimes also called "positive demand externalities") can be found in the literature. Basically, it can be stated that network externalities exist if the utility of a product for a customer depends on the number of other customers who have also bought and use this product. These network externalities can be indirect or direct, whereby we concentrate on the latter. A typical example for a product with direct network effects is the telephone. The utility that a telephone system can generate increases with the amount of feasible communication connections. Other examples are e-mail, fax, instant messaging, etc. each of which satisfies communication needs. But none of these products can generate any utility for a user on its own. In order to create utility the existence of other users adopting this product is required. Accordingly, the product's utility changes dynamically with the number of users, i. e., the installed base.

The installed base B_t determines the utility of products influenced by network externalities. In terms of direct network externalities the utility is based on B_t exclusively since utility can only be generated by interconnections within the underlying network solely. Accordingly, the utility of a product with direct network externalities is a function of the number of feasible connections I_t. The number of connections is determined by the number of users on the one hand and the technological restriction of the network on the other hand, whereby the latter one represents the number of users being able to communicate via the network simultaneously (for instance, classical telephone system $r = 2$, telephone conferencing $r \leq 2$). Thus, U_t can be calculated by the formula:

$$U_t = U_t(I_t) = \sum_{k=2}^{n} \binom{B_t}{r} = \sum_{k=2}^{n} \frac{B_t!}{r!\,(B_t - r)!}\,, \quad \text{whereby } r,\ B_t > 0\,. \qquad (15)$$

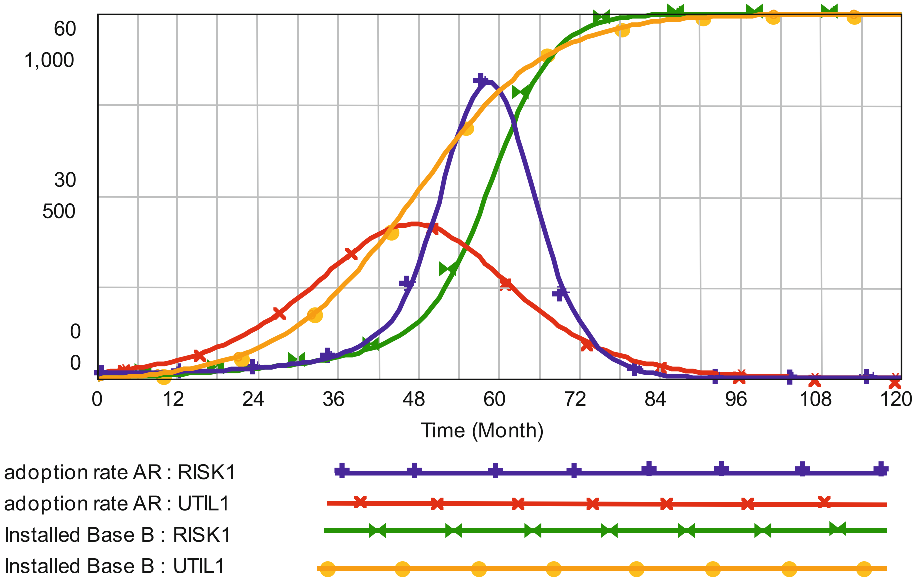
Since the achievable utility of a product with direct network externalities depends exclusively on the network size the adoption process depends on the decision of potential users influencing the diffusion process significantly. This leads to two different effects: Firstly, the utility for each user grows exponentially with an increasing amount of actual users according to the formula. This implies that the more people are attracted the more are part of the network leading to an exponential growth of the diffusion process which is referred to as the "bandwagon effect": The higher the number of people the more are decoyed as well resulting in a reinforcing process. Although this effect occurs with conventional products as well, in case of products with direct network externalities, it is much stronger since the exponentially growing utility has a greater impact on the diffusion process.

Secondly, utility must be created by the utilization of the users first in order to establish a new communication network which determines the diffusion process significantly since products influenced by direct network externalities cannot generate an original utility by itself. Accordingly, early adopters of a network are confronted with the risk that they cannot derive sufficient benefit from the network so that they must rely on potential users to follow entering the network in the future. Therefore, the adoption decision of a potential user depends on the future decision of other potential users. All in all, this leads to a hesitating behavior of all potential adopters resulting in an imaginary network barrier, which is based on the risk of backing the wrong horse, which is also known as the "penguin effect".

Finally, another important aspect must be considered when analyzing the diffusion process of products with network externalities. In terms of conventional products the decision to buy a product is the final element of the decision process. Contrary to that, concerning products with network externalities the adoption process is not finished with the decision to enter a network since the subsequent utilization of the product is important for the diffusion process. If the expected utility of the communication product cannot be achieved users may stop using this product leading to a smaller installed base and a lower network utility which is important for other users that may stop their utilization as well and for the adoption process of potential users.

In the following, the basic structure of the underlying model will be described. Analogously to the model presented in the preceding paragraphs there exists a group of potential users (in this model, we only focus on the core diffusion process without considering competitors or latent demand in order to keep the complexity of the model low.) If these potential users decide to adopt the communication product, they become part of the installed base *B*. The adoption process is illustrated by the adoption rate *AR*, which is primarily influenced by the variable *word of mouth*. In order to consider the average utility per user – as it is necessary for analyzing products with network externalities – the imitation coefficient β has been endogenized contrary to the classical Bass model. Therefore, the variable β is influenced by the "word-of-mouth" effect which depends on the average utility per user. If actual utility is bigger than the desired utility all individuals in contact with users adopt and buy the product. If it is smaller, however, only a fraction adopts. The size of this fraction depends on the distance between actual and desired utility.

Figure 9 depicts two simulation runs showing the system behavior of the diffusion process of conventional products and products influenced by direct network externalities. Graph UTIL1 represents the *adoption rate*, i. e., the amount of buyers of a conventional product per period. The graph UTIL1 shows the behavior of the variable *installed base B* which is the accumulation of *adoption rate* (note that the graphs have a different scale). The graphs



Diffusion of Innovations, System Dynamics Analysis of the, Figure 9
Comparison of diffusion behavior

RISK1 show the system behavior for products influenced by direct network externalities, i. e., the *adoption rate AR* and the corresponding *installed base B*.

A comparison of both simulation runs shows that diffusion needs longer to take off in terms of products influenced by direct network externalities, but showing a steeper proceeding of *Installed Base B* in later periods. This behavior can be verified comparing the adoption rates of the two runs: although adoption starts later with an endogenously generated adoption fraction, it nevertheless has a higher amplitude. This behavior can be interpreted as the penguin effect and the bandwagon effect.

Finally, it has to be taken into account that some users of the installed base might quit to use the product since they are disappointed from its utility. Accordingly, it is an important issue to find ways in order to raise the patience of users to stay within the network. That gives potential users the chance to follow into the network which will increase the utility for the user as well.

From the simulation analysis the following conclusions can be drawn. The importance of the installed base for a success diffusion process is shown. Without a sufficient amount of users it is not possible to generate a utility on a satisfying level which prevents potential users to enter the network or even making users leave the network. Accordingly, ways must be found to increase the utility that a network creates for a user in order to reach the critical mass. This can be done in several ways of which some will be discussed briefly. One possible way is to increase the installed base by compatibility to other networks. Furthermore, the risk to back the wrong horse can be mitigated by product pre-announcements in order to lower the imaginary network barrier by making potential users familiar with the product. Another possibility is to increase the group of relevant users, i. e., to enlarge the average group size within the network, since not all users are equally important for a potential user. Furthermore, the technological potential can be improved by introducing multilateral interconnections between the members of a network.

Representing Managerial Decision Making in Innovation Diffusion Models

Subsequently the basic structures of innovation diffusion processes described above will be extended and simulated to demonstrate the impact of managerial decision-making on the diffusion of innovations. The model used for the simulations serves as a simulator to determine how individual strategies can accelerate or hamper market penetration and profit performance. The models are not designed to predict the basic market success or failure of innovations. Although, they are rather comprehensive, several assumptions apply here as for all models. E.g., in all model runs, the basic market acceptance of the innovation is assumed. Furthermore, the simulations assume for the moment that no competition exists.

Dynamic Pricing Without Direct Competition

In a first step the basic model from Subsect. "Base Structure of a System Dynamics-Based Model of Innovation Diffusion" is extended to generate dynamic cost behavior as suggested in Fig. 5. Standard costs are the basis for the calculation of prices – an important decision variable. Experience curve effects are modeled based on cumulated production in order to map the long-term behavior of standard cost. The actual costs of a product in a certain period are derived from the standard cost modified for variations resulting from capacity utilization.

The concept of experience curve effects suggests a direct relationship between cumulated production $X_{(t)}$ and average standard cost per unit $c^{s}_{(t)}$, adjusted for inflation; where c^{s} defines standard unit cost at the planned level of production. Every doubling of $X_{(t)}$ is associated with a cost reduction in real terms by a constant percentage according to:

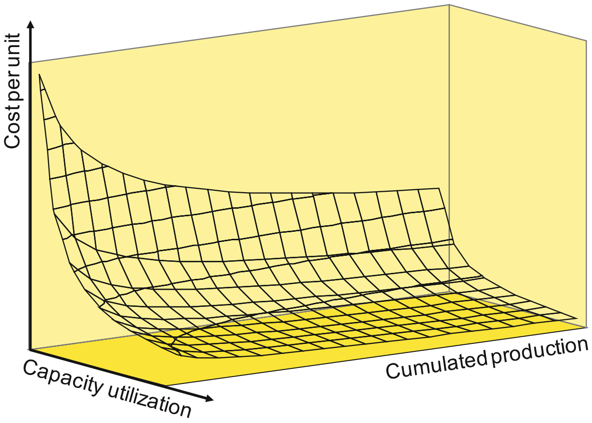
$$c^{s}_{(t)} = c^{n}\left(\frac{X_{(t)}}{n}\right)^{-\delta} \tag{16}$$

where c^{n} stands for the cost of unit n ($n \subseteq X$) and δ represents a constant depending on the experience rate. For many businesses experience rates of 10% to 20% have been observed and ample empirical evidence for this relationship is available.

The costs of a product in each period of time $C_{(t)}$ are a function of cumulated production $X_{(t)}$ and capacity utilization determined by the production volume of a period $x_{(t)}$ as defined in Eq. (17). Figure 10 shows the behavior of the dynamic cost function

$$C_{(t)} = \Phi\left(X_{(t)}, x_{(t)}\right) . \tag{17}$$

Furthermore, the model comprises elements of (i) market development, (ii) product pricing and its impact on the profits from producing and selling the products, i. e., the operating results, and (iii) resource allocation, e. g., capital investment, production volume, and quality control. Pricing and quality affects the coefficients of innovation α and imitation β from Eq. (10). Figure 11 shows the run of a model version including market development.



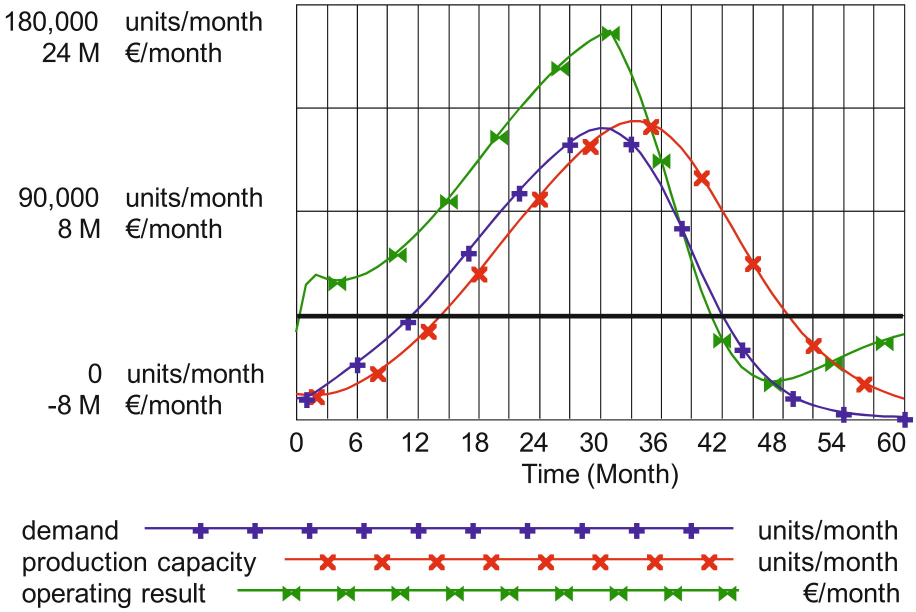
Diffusion of Innovations, System Dynamics Analysis of the, Figure 10
Dynamic cost function

The time behavior of production, demand, and operating results duplicate usual characteristics of the life cycle of a successful innovation. After the product launch, additional customers can be gained from an untapped market as diffusion and thereby product awareness proceeds and prices decline. The maximum of demand from imitators – the quantitatively most important fraction of demand – is reached when the amount of possible communications between potential customers and adopters reaches its maximum. The decreasing level of potential customers and the depletion of the untapped market cause the decline towards the end of the simulation. The behavior also shows that demand rises much faster than the company can increase its production capacity. The behavior of Fig. 11 will serve as reference mode for further analysis.

Pricing strategies and decisions are additional important elements, which require an extension of the model. The problem of the "right price" for a new product is essential but still unsolved in the area of innovation management. Difficulties to recommend the optimal pricing policy derive in particular from the dynamics in demand interrelations, cost development, potential competition, and the risk of substitution through more advanced products. Regardless of this complex framework, several attempts in management science try to derive and to apply optimal pricing policies. However, they are faced with difficulties, both mathematical and practical. Their results are too complicated to support actual pricing decisions. Therefore simulation studies found more frequently their way into management science.

The extended model includes four predefined pricing policies to investigate their impact on market development on operating results:

Myopic profit maximization assuming perfect information about cost and demand. The optimal price p^{opt} is derived from elasticity of demand $\varepsilon_{(t)}$ and per unit standard



Diffusion of Innovations, System Dynamics Analysis of the, Figure 11
Reference mode of the basic innovation diffusion model

cost $c^{s}_{(t)}$ considering the impact of short term capacity utilization:

$$p^{\mathrm{opt}}_{(t)} = c^{s}_{(t)} \cdot \frac{\varepsilon_t}{\varepsilon_t - 1} \,. \tag{18}$$

Skimming price strategy aims at serving innovative customers with high reservation prices and then subsequently reduces prices. The model applies a simple decision rule modifying $p^{\mathrm{opt}}_{(t)}$ through an exponential function that raises the price during the first periods after market introduction:

$$p^{\mathrm{skim}}_{(t)} = p^{\mathrm{opt}}_{(t)} \cdot \left(1 + a \cdot \mathrm{e}^{\frac{-t}{T}}\right) \,. \tag{19}$$

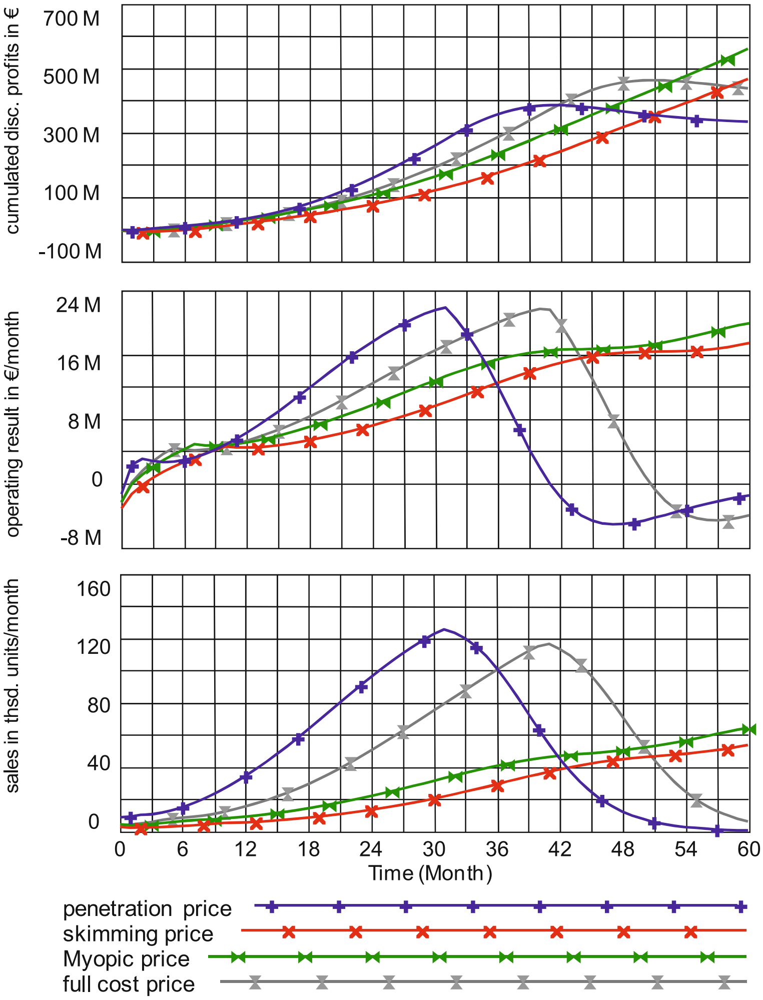
Full cost coverage, i. e., standard cost per unit plus a profit margin π to assure prices above cost level even during the early stages of the life cycle:

$$p^{\mathrm{fcc}}_{(t)} = c^{s}_{(t)} \cdot \pi \,. \tag{20}$$

Penetration pricing aims at rapidly reaching high production volumes to benefit from the experience curve and to increase the number of adopters. It uses a similar policy as for the skimming price, but instead of a surcharge it decreases prices early after market introduction:

$$p^{\mathrm{pen}}_{(t)} = c^{s}_{(t)} \cdot \pi \cdot \left(1 - a \cdot \mathrm{e}^{\frac{-t}{T}}\right) \,. \tag{21}$$

The simulation runs shown in Fig. 12 give an overview of the development of profits, cumulated profits, and sales for the four pricing strategies discussed above. The model assumes the following: (1) there is an inflow from the un-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 12
Comparison of the outcome of pricing strategies

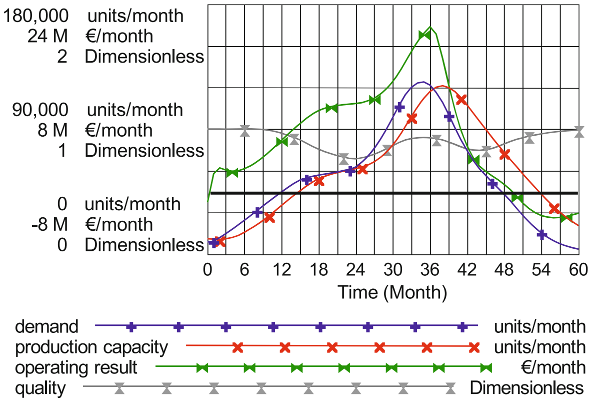
tapped market, which depends on the dynamic development of prices; (2) there is no risk of competition; (3) repeat purchases do not occur. Taking profits into account, Fig. 12 indicates that – over the time horizon observed –, the classic pricing rule of profit optimization leads to superior results from a financial point of view. However, if judged by the market development, the strategy of penetration prices is the appropriate strategy. This strategy allows rapid penetration of the market by setting relatively low prices, especially in the early stages of the life cycle. The combined price and diffusion effects stimulate demand and reduce the risk of losing potential customers to upcoming substitution products.

Figure 12 also indicates a disadvantage of the penetration strategy. Since the market is already completely satisfied after period 54, there is only little time to develop and introduce a new product in the market successfully. The slower market growth of the skimming and optimum price strategy leaves more time for the development of a new product, but the attractive profit situation and the slow development also increase the risk that competitors might enter the market. In a dynamic demand situation where prices influence market growth, where substitution or competition can occur, and where delivery delays eventually accelerate the decision of potential buyers to turn to other products, a strategy of rapid market penetration seems to be the most promising one. It will, therefore, be the basis for the following simulation runs investigating manufacturing's role in innovation management.

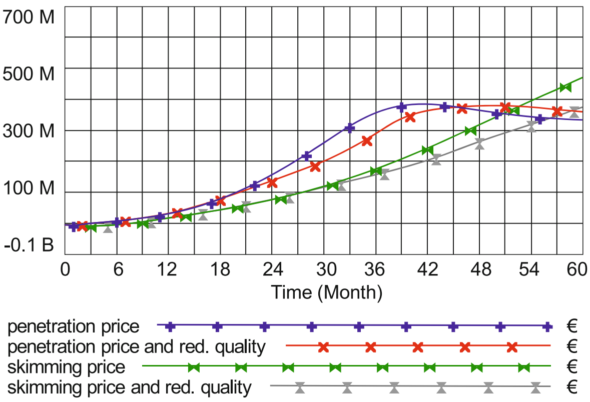
Linking Manufacturing-Related Decision Variables

The role of manufacturing is important for the successful management of innovations. Manufacturing has to provide sufficient capacity to produce the goods sold. The investments to adjust capacity influence a company's ability to meet demand and deliver on time. It is assumed that the necessary financial resources for the investments are available. The aggregated capacity provided by the company includes both, machinery equipment and production personnel. Since the manufacturing function also has to ensure the quality of the output through dedicating a portion of its total available capacity to quality control, the capacity resources can be used to either manufacture the products or to assure the desired level of quality. Capacity allocation to improve quality takes away capacity for production. This additional feedback structure – as indicated in Fig. 5 – maps the allocation of resources for quality control to the achieved ability to meet product demand. If manufacturing capacity does not meet demand, a temporary reduction of capacity for quality assurance seems a plausible strategy. Quality control resources than are allocated to manufacturing rather than testing whether quality standards are met. In this scenario, it would be expected that total cost remain unchanged and the additional manufacturing capacity gained through the reallocation can be used to provide more products to the customers, increase sales, and improve the overall results.

Figure 13 shows the simulation assuming the same scenario as in the base mode together with penetration prices and reduced quality resources if demand exceeds production capacity. It also shows a quality index plotted as an additional variable. Quality is defined to be 1, if the actual quality capacity equals a standard value of quality resources necessary. It is assumed that 10% of total production capacity is necessary to assure 100% quality. For values above the 10%-level, quality is better; for values below, it is poorer. The simulation indicates that the policy of reduced quality resources successfully decreases the discrepancy between demand and production as seen in the reference mode of Fig. 11. This results from the increased proportion of capacity used for production and an additional effect caused by lower product quality, which then decreases demand. Although the maximum sales are nearly the same in the simulation of reduced quality control strategy, the peek demand occurs around 5 months later. Instead of gaining higher sales only the shape of the life cycle changed. However, operating results had improved, in particular the sharp decline of profits in the base mode of the simulation could be slowed down and losses could be avoided. The reduced quality control strategy caused a slower capacity build-up and therefore, when product sales declined capacity adjustment was easier to achieve.



Diffusion of Innovations, System Dynamics Analysis of the, Figure 13
Reduced quality control



Diffusion of Innovations, System Dynamics Analysis of the, Figure 14
Cumulated discounted profits – penetration vs. skimming pricing in combination with quality control strategies

From the financial point of view the strategies of penetration prices and reduced quality control fit quiet well.

The results are different if a strategy of quality reduction is used in combination with a strategy of skimming pricing. Figure 14 compares the outcome of cumulated discounted profits for the strategy of reduced quality and penetration prices or skimming prices with the development of the reference mode – the simulations without quality adjustment. The behavior indicates that in the case of skimming prices, quality reductions slow down the development of the market and cumulated profits significantly.

The simulation results raise the question whether emphasizing quality when demand is higher than capacity would be a more appropriate way to react. As the upper part of Fig. 15 points out, the strategy of emphasized quality leads to an accelerated product life cycle in the case of the penetration pricing strategy. Tremendous capacity build-up is necessary after the introduction of the new product. As demand declines, a plenty of capacity is idle, causing significant losses during the downswing of the product life cycle.

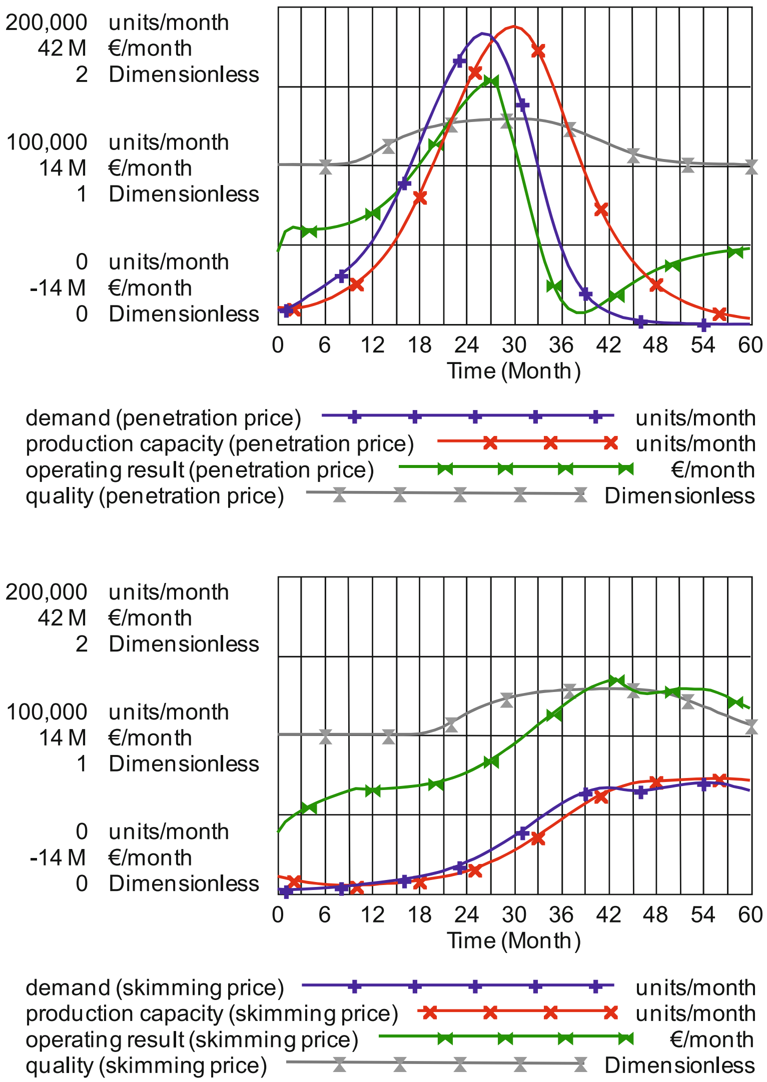
Emphasizing quality turns out to be more effective in the case of skimming prices. The additional demand gained from quality improvements also accelerates the product life cycle, but at a much slower rate and leads to improved cumulated profits. Emphasizing quality in combination with skimming or optimum prices leads to improved cumulated profits, compared to both, the simulation without quality reaction and the quality reduction run.

The simulations show the importance of a detailed judgment of strategic choices. Strategies must be consistent with each other and with the real world structures mapped by the model. The simulations above assume a situation without existing or potential competition. In such an environment there is no interest paid for fast market penetration. Hence, a penetration pricing strategy is the most unfavorable alternative. However, this changes if structural elements are added to the model that incorporate competition – even as in the simple structure from Fig. 5, which considers the loss of demand to a competitor. Lost demand therefore is represented as a process equivalent to the imitative demand from Eq. (9). The calculation of lost demand starts in period 15 through an initial switching of a potential customer to the competitor. This switch starts a process that drives demand for the competitors' products and is influenced through the quality the company offers. If the company provides poor quality, more potential customers and market potential from the untapped market will directly move to the competitor. The accumulation of lost demand corresponds to the number of adopters the competitors gained over time. Simulations with these additional structures give some additional insights (Fig. 16).

Penetration pricing leads again to the fastest market development. In the competitive surrounding, however, emphasizing quality accelerates the market development and leads to better performance than quality reductions. This is in contrast to the simulations without competition shown in Fig. 13 to Fig. 15. Skimming prices in combination with reduced quality control shows the poorest financial and market performance. A strategy of reduced quality control causes in the competitive environment the demand to increase at a slower rate than in the base run, where no quality adjustments were made when demand exceeded capacity. In both cases, the skimming and the penetration price scenario, quality reductions lead to the poorest performance.

Linking R&D and New Product Development

The models discussed above are able to generate under different conditions the typical diffusion patterns of new products in the market place. However, these models do not consider the stage of new product development. New products have to be developed before they can be introduced into the market. A costly, lengthy, and risky period of R&D has to be passed successfully. The diverging trends of shortening product life cycles and increasing R&D costs show the importance of an integrated view of all innovation stages. In the remainder, a comprehensive model comprising both, the process of R&D and an oligopolistic innovation diffusion with subsequent product generations



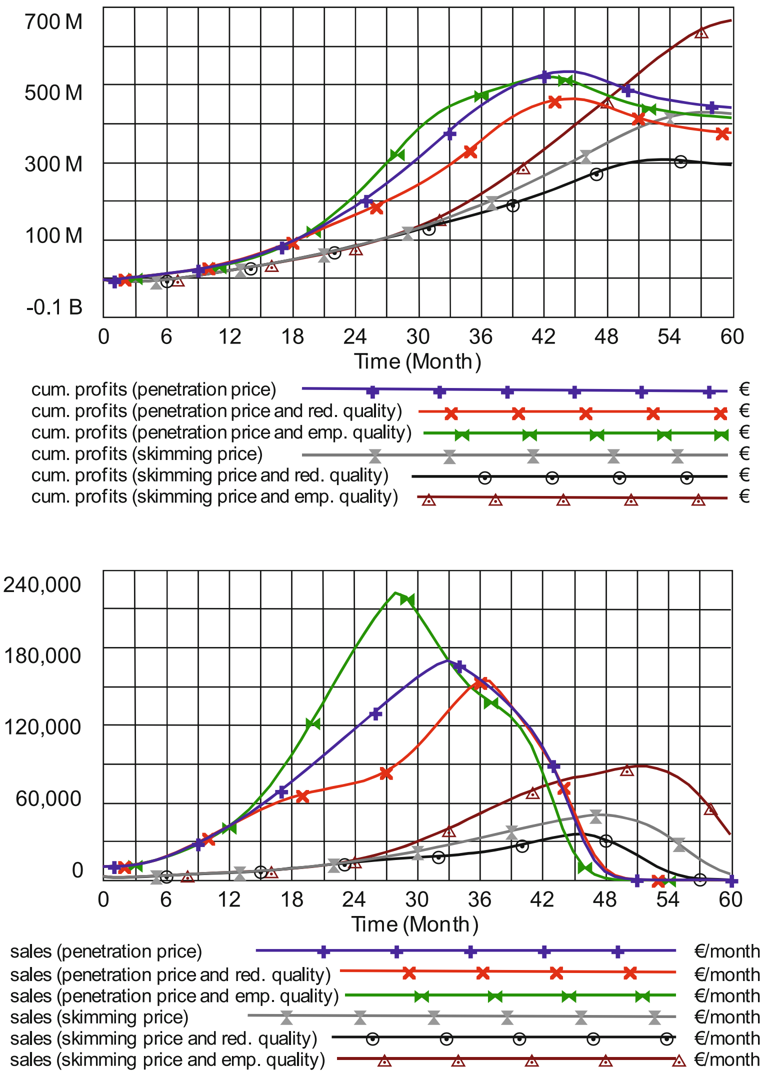
Diffusion of Innovations, System Dynamics Analysis of the, Figure 15
Emphasized quality in all innovation stages

is used to investigate the interrelations between the stages of innovation processes. The integration of both modules is shown in Fig. 17.

The volume and the intensity of the research and development activities feed the R&D-process. The number of research personnel determines the volume. Since R&D personnel requires resources like laboratory equipment, material for experiments etc., the intensity of R&D depends on the budget available for each person working in the R&D sector. This information is calculated in a more comprehensive model in the sector of R&D planning, which also includes policies about resource allocation within the research and development stages, i. e., mainly the question of how much to spend on which new product development project.

Depending on the volume and the intensity of R&D, the technological knowledge of each product generation for each company evolves over time. The module of the R&D-process feeds back the current state of the technological knowledge for each company and product generation.

The basic assumptions of the model are as follows. The model maps the structures of two competitors. Both competitors can introduce up to five successive product generations. The initial values of the model ensure that all competitors start from the same point. All firms have already

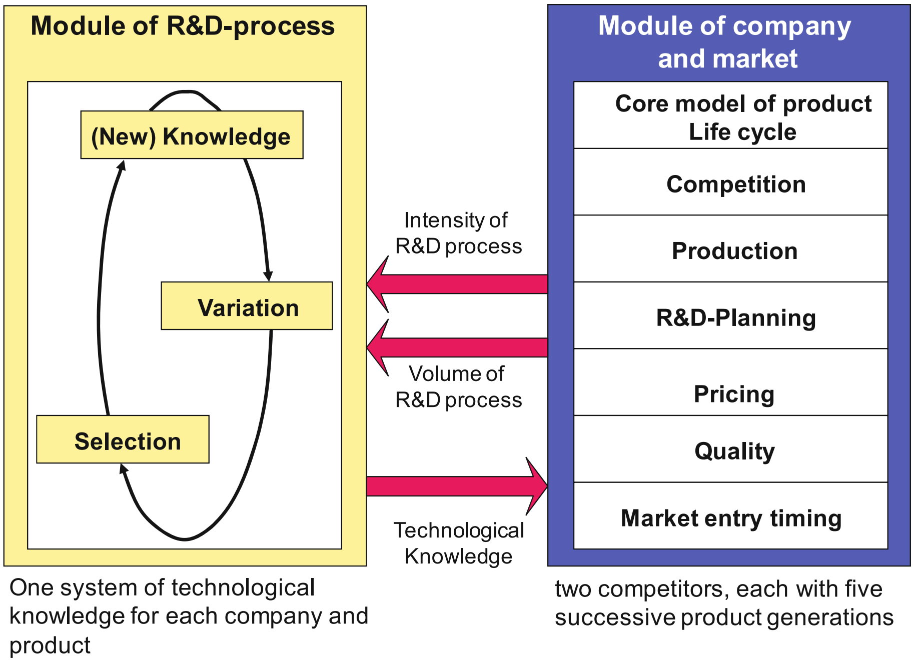


Diffusion of Innovations, System Dynamics Analysis of the, Figure 16
Behavior of the base model including simple competitive structures

introduced the first product generation and share the market equally. The resources generated by the first product are used to develop subsequent product generations. In the base run each company follows the same set of strategies. Therefore, except for minor differences resulting from the stochastic nature of the R&D-process, they show the same behavior over time. Figure 18 provides a simulation run of the model with all modules and sectors coupled.

The curves show for a single company the development of the sales of the products and the total sales. They emphasize the importance of a steady flow of new and improved products. Without on-time replacement of older products, the total sales of the products will flatten or deteriorate like in the simulation around periods 44, 92, and 116. The model also generates the typical s-shaped curves of technological development (lower part of Fig. 18). Each product generation has a higher technological potential and the knowledge developed for the preceding product generations partly can be used by the successive product generations. For this reason the subsequent product generations start at a level different from zero.

In a dynamic environment such as the computer industry, where investments in R&D and manufacturing equipment are high, the product life cycles are short, and time-to-market as well as time-to-volume are essential variables, it is important to understand the dynamic con-


Module of R&D-process
(New) Knowledge
Variation
Selection
One system of technological knowledge for each company and product
Intensity of R&D process
Volume of R&D process
Technological Knowledge
Module of company and market
Core model of product Life cycle
Competition
Production
R&D-Planning
Pricing
Quality
Market entry timing
two competitors, each with five successive product generations

Diffusion of Innovations, System Dynamics Analysis of the, Figure 17
Linking R&D-processes with corporate and market structures

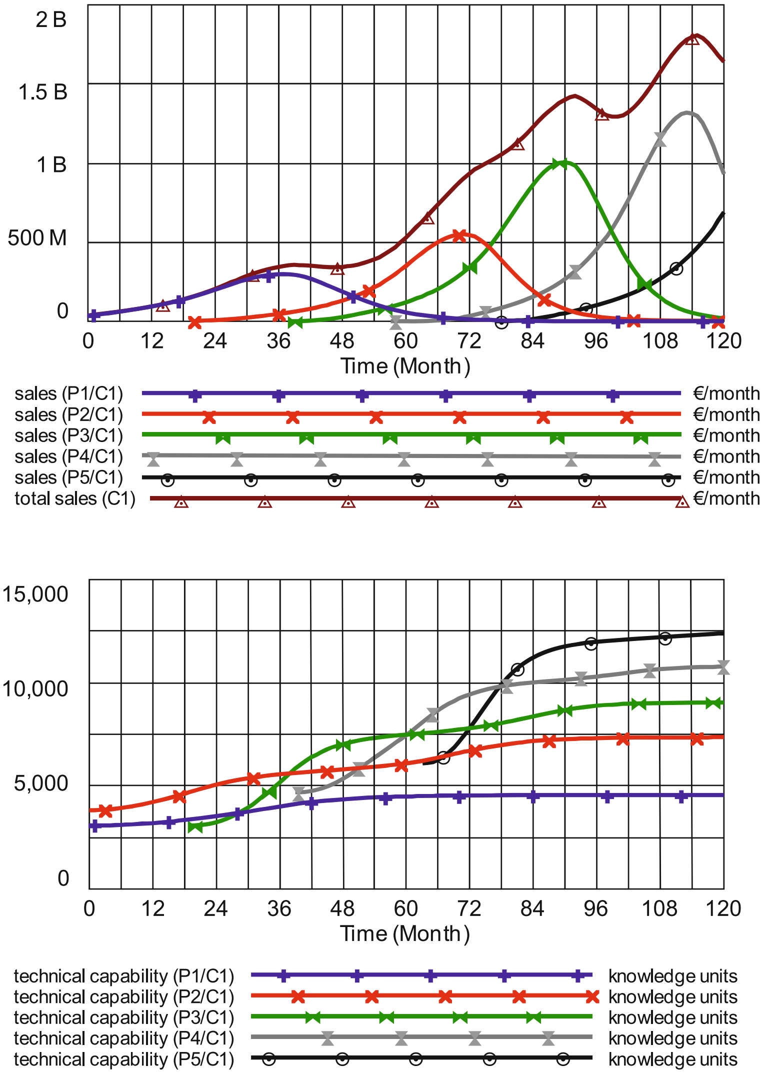
sequences of decisions and strategies in the different areas. Figure 19 describes some of the important feedback loops linking the process of invention to the processes of innovation and diffusion.

Central element in the figure is the calculation of the sales of a company according to Eqs. (11) and (14). The coefficients of innovation and imitation are influenced by the multiplier of relative competitive advantage, which depends on the relative technical capability and the price advantage of a company. The technical capability of the products is influenced by the strength of its R&D-processes and the total amount of R&D expenditures. Empirical studies in Germany have shown that measures like sales volume, profits or R&D budgets of earlier periods are quite common as a basis for R&D budgeting. However, using historic sales volume as a basis to determine R&D budgets invokes the positive feedback loop "competing by technical capability". With an increasing number of products sold and growing value of sales the budget and the number of personnel for R&D grow. This leads to an improved competitive position, if the technical capabilities of a product increases. The higher the sales volume, the better is the resulting competitive position. This produces increasing coefficients of innovation and imitation and leads to higher sales. This budgeting strategy is implemented in the model for the next simulation runs.

The second loop "price competition" links pricing strategies to sales volume. The actual price of a product is influenced by three factors. The first factor, standard costs, is endogenous. As cumulated production increases, the experience gained from manufacturing causes declining standard costs. The second and third elements influencing the calculation of prices are exogenous: parameters which define the pricing strategy and demand elasticity. Caused by increasing cumulated production, standard costs fall over the life cycle and prices are also declining. Lower prices affect the relative price and improve the effect of price on the coefficients of innovation and imitation, which leads to increased sales and higher cumulated production.

The loop "pricing limits" reduces the effects of the reinforcing loops described above to some extent. The standard cost and price reductions induce – ceteris paribus – a decrease in the sales volume and set off all the consequences on the R&D-process, the technical know-how, the market entry time and sales shown in the first feedback loop – but in the opposite direction. Additionally, since standard cost cannot be reduced endlessly this feedback loop will show a goal seeking behavior.

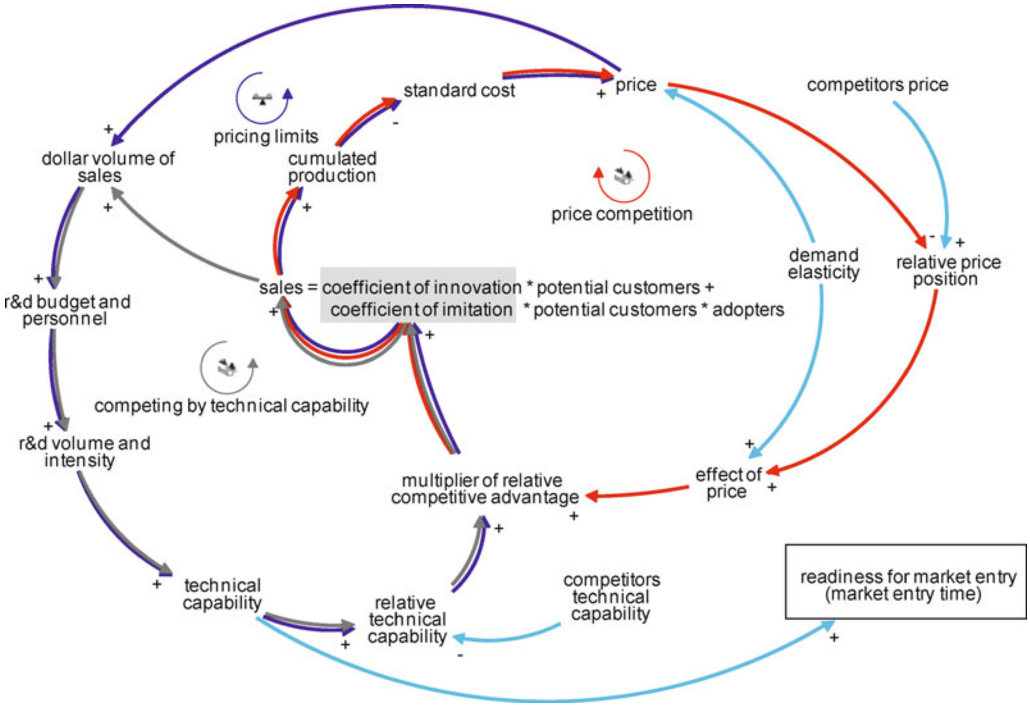
With equivalent initial situations and the same set of strategies, both companies behave in an identical way for all product generations. If one company has a competitive



Diffusion of Innovations, System Dynamics Analysis of the, Figure 18
Exemplary behavior of the integrated innovation model

advantage, the reinforcing feedback loops suggest that this company will achieve a dominating position. In the simulation shown in Fig. 20, both competitors have the same competitive position for the first product generation. But the first company will be able to enter the market 2 months earlier than the competitor, because the initial outcome of the R&D process is slightly better than the second company's second product generation. Both competitors follow a strategy of skimming prices and demand elasticity has the value −2.

The initial gain in the outcome of the R&D-process initiates a process of sustained and continuing competitive advantage for the first company. It will improve continuously, since the positive feedback loop "competing by technical capability" dominates. The first company's advantage in the market introduction leads to an increasing readiness for market entry. It is able to launch the third product generation 3 months earlier than the follower and will introduce the fourth product generation in period 112. The follower is not able to introduce its fourth generation during the time horizon of the simulation, i. e., the pioneers advantage has extended to more than 8 months. The first company's competitive advantage is a result of the slightly higher initialization of the knowledge system and the dominance of the positive feedback loops, which causes shortened time-to-market and higher sales volume


standard cost
price
competitors price
pricing limits
cumulated production
dollar volume of sales
price competition
demand elasticity
relative price position
r&d budget and personnel
sales = coefficient of innovation * potential customers + coefficient of imitation * potential customers * adopters
competing by technical capability
r&d volume and intensity
multiplier of relative competitive advantage
effect of price
technical capability
relative technical capability
competitors technical capability
readiness for market entry (market entry time)

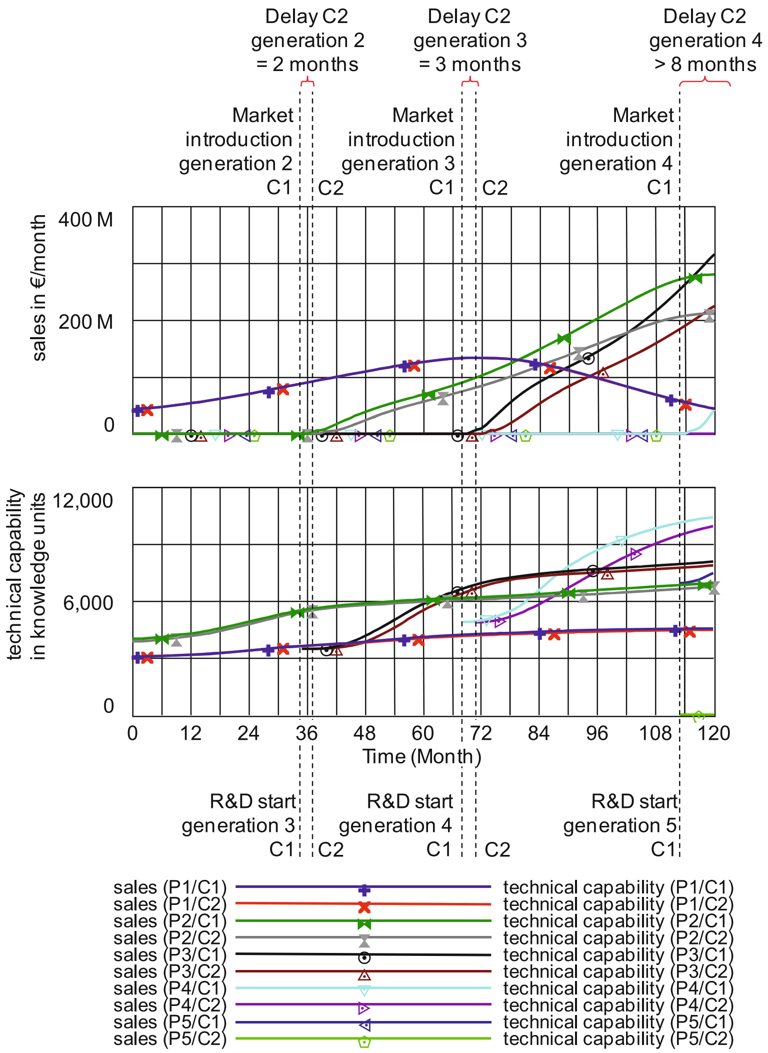
Diffusion of Innovations, System Dynamics Analysis of the, Figure 19
Feedback structure influencing the diffusion process

over all successive product life cycles. Additionally, the technical capabilities of both competitors' product generations show the same reinforcing effect. The difference between the technical capability of both competitors increases in favor of company 1 until they approach the boundaries of the technology.

Although literature discusses a variety of models to find optimal pricing strategies, these models usually only consider the market stage of a new product and neglect the interactions with the development stage of a new product. Pricing decisions not only drive the diffusion of an innovation, but they also have a strong impact on the resources available for research and development. Since the comprehensive innovation model links the stages of developing and introducing a new product, the following simulations will show the impact of pricing strategies on performance in a competitive environment. In the analysis shown the first company uses the strategy of skimming price for all product generations. The second company alternatively uses a skimming price strategy in the first model run, myopic profit maximization strategy in the second run, and the strategy of penetration prices in the third run. The initial conditions are identical, except the price strategy settings. Sales volume, market position, and cumulated discounted profits are used to judge the advantages of the alternative pricing strategies. Market entry time is used as a measure of time-to-market.

The logic behind the skimming price strategy is to sell new products with high profit margins in the beginning of a life cycle to receive high returns on investment, achieve short pay off periods, and high resources for the R&D-process. However, in a dynamic competitive setting the strategy of myopic profit maximization and penetration prices achieve better results (Fig. 21). Company 1 which uses a skimming price strategy achieves the lowest sales volume. Myopic profit maximization prices and penetration prices of the second competitor causes the sales to increase stronger through the combined price and diffusion effect.

The results are confirmed if the variable market position – an aggregate of the market share a company has for its different products – is taken into account. For values greater than 1 the market position is better than the one of the competitor. Using the penetration strategy, company 2 can improve its market share, achieve higher sales volume and therefore has more resources available for R&D. This enables it to launch new products earlier than company 1. As shown in Table 1, the advantage of time-to-market increases from product generation to product generation.



Diffusion of Innovations, System Dynamics Analysis of the, Figure 20
Reinforcing effects of initial competitive advantage

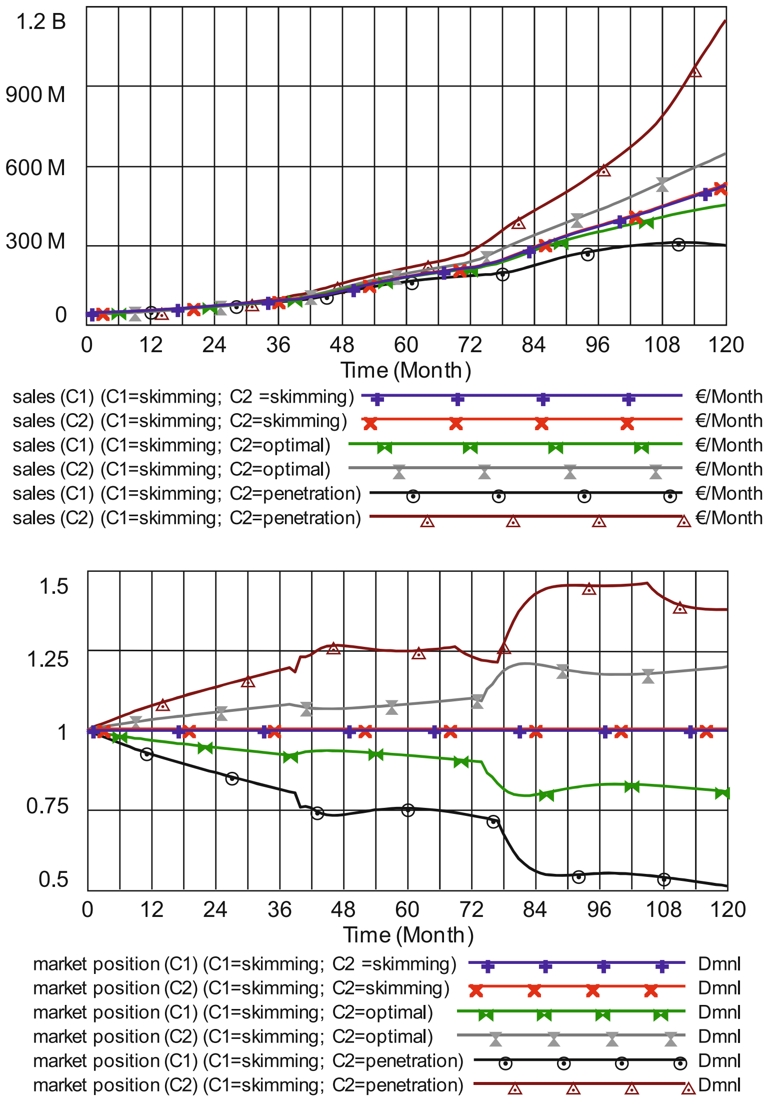
Diffusion of Innovations, System Dynamics Analysis of the, Table 1
Consequences of pricing strategies on market entry time

	Product generation 2			Product generation 3			Product generation 4		
Pricing strategy C2*	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2
Skimming prices	38	38	0	71	71	0	n.i.	n.i.	–
Profit maximization	36	35	1	71	69	2	n.i.	118	> 2
Penetration prices	37	35	2	74	66	8	n.i.	102	> 8

*C1 uses skimming prices in all simulations; **n.i. = product was not introduced

The improvement in time-to-market for the first company's second product generation results from the slightly higher sales volume compared to the use of skimming pricing strategies for both competitors. The second company achieves the strongest improvements in time-to-market if it uses a penetration pricing strategy.

In terms of cumulative profits (Fig. 22) one would expect that skimming prices should generate the highest



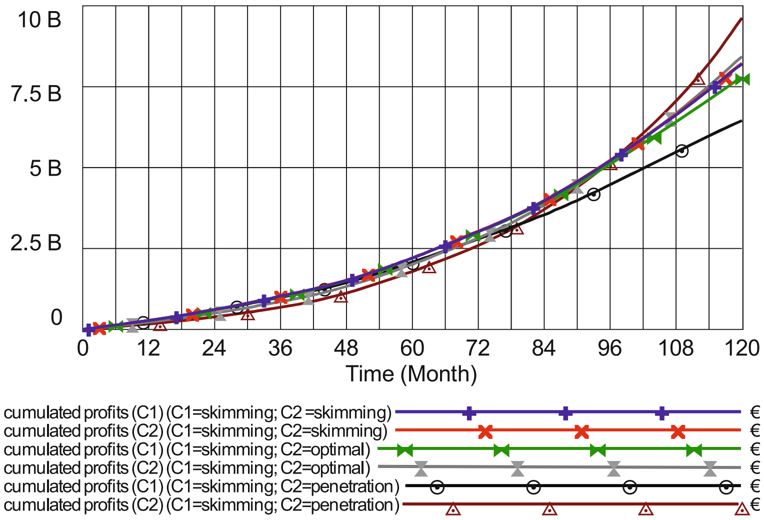
Diffusion of Innovations, System Dynamics Analysis of the, Figure 21
Sales volume and market position for different pricing strategies

cumulated profits, however, this is not true. Penetration prices generate the highest results followed by skimming prices. The strategy of myopic profit maximization shows the least favorable outcome.

The simulations so far assumed a price response function with a constant price elasticity ε of -2. Since price elasticity influences both, the demand for a product as well as the price level (cf. Fig. 19), the influence of price elasticities have to be investigated before recommendations can be made. Assuming that company 1 uses a strategy of skimming prices and the second competitor follows a strategy of penetration pricing, Fig. 23 shows the time path of cumulated discounted profits and market position for ε between -3.2 and -1.2.

Due to the different profit margins – resulting from myopic profit maximization being the basis for price calculation – the use of the absolute value of the cumulated profits is not appropriate. Therefore, the second company's share of the total cumulated profits is used for evaluation purposes. The measure is calculated as

$$\left(\frac{\text{cum.profits}_2}{\sum_{i=1}^{2} \text{cum.profits}_i}\right) .$$



Diffusion of Innovations, System Dynamics Analysis of the, Figure 22
Time path of cumulated profits

The first graph in Fig. 23 shows that the initial disadvantage of the second company rises with increasing demand elasticity. However, its chance of gaining an advantage increases as well. In the case of lower demand elasticities ($\varepsilon > -1.7$) firm 2 cannot make up the initial disadvantage during the whole simulation. For demand elasticities ($\varepsilon > -1.4$) the cumulated profits ratio even deteriorates. Considering the market position the picture is similar. For demand elasticities $\varepsilon > -1.6$ the penetrations strategy leads to a loss in the market position in the long run. The improvements resulting from the introduction of the successive product generations are only temporary.

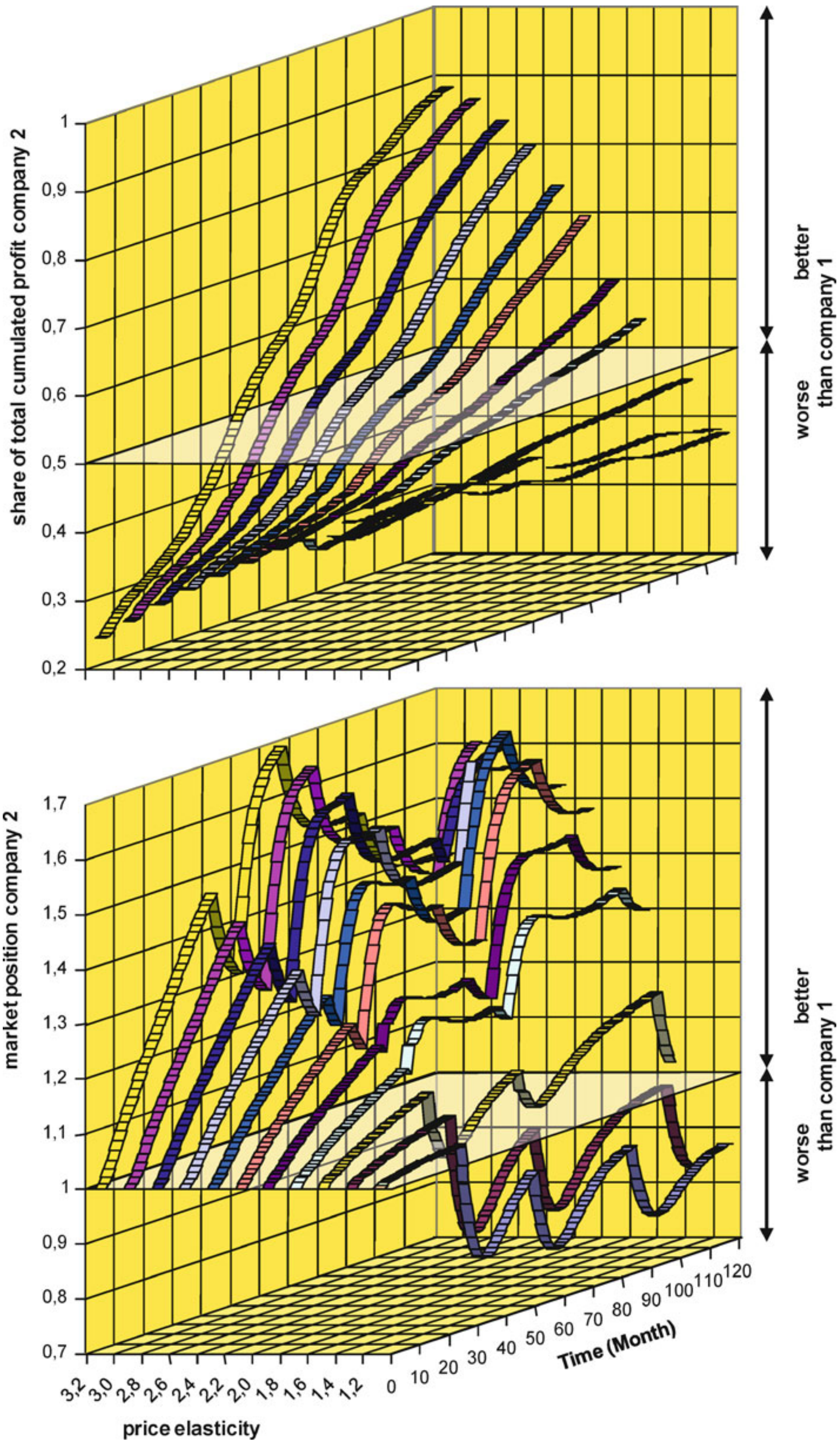
Managerial Implications

The simulations above lead to the insight that general recommendations for strategies are not feasible in such complex and dynamic environments. The specific structures like competitive situation, demand elasticity, or strategies followed by the competitors have to be taken into account. Recommendations only can be given in the context of the specific situation. Furthermore, the evaluation of strategies depends on the objectives of a company. If a firm wants to enhance its sales volume or the market share, the strategy of penetration pricing is the superior one. Viewing cumulative profits and the readiness for market entry as prime objectives, the strategy of skimming prices is the best. However, these recommendations hold only for high demand elasticities. Furthermore, the model does not consider price reactions of competitors. The evaluation of improved strategic behavior would become even more difficult. The outcome and the choice of a particular strategy depend on many factors that influence the diffusion process. The dynamics and the complexity of the structures make it almost unfeasible to find optimal solutions. Improvements of the system behavior gained through a better understanding, even if they are incremental, are steps into the right direction.

Future Directions

The series of models presented here are designed in a modular fashion. They offer the flexibility to be adapted to different types of innovations, to different structures, initial conditions and situations. The models provide the opportunity to investigate courses of action in the setting of a management laboratory. They allow one to investigate different strategies and to learn in a virtual reality. They emphasize the process of learning in developing a strategy rather than the final result. To support learning processes, the models could be combined with an easy-to-use interface and serve as a management flight simulator which allows one to gain experience and understanding from playing.

Although the models cover a variety of different aspects in the management of innovations, they still can be extended. Besides more detailed mapping of corporate structures behind managerial decision processes the structures representing the diffusion process can be extended in various ways. Although some research already discusses the problems of mapping the substitution among succes-



Diffusion of Innovations, System Dynamics Analysis of the, Figure 23
Impact of demand elasticity on performance measures

sive product generations, this area deserves further attention. In particular in high-tech industries with short product life cycles the interrelations between successive product generations strongly influence the overall success of a company. Furthermore, the diffusion structures could be extended to include cross-buying and up-buying behavior of customers and by that link models of innovation diffusion to the field of customer equity marketing.

Bibliography

Primary Literature

1. Bass FM (1969) A New Product Growth Model for Consumer Durables. Manag Sci 15:215–227
2. Bental B, Spiegel M (1995) Network Competition, Product Quality, and Market Coverage in the presence of network externalities. J Ind Econ 43(2):197–208
3. Boston Consulting Group (1972) Perspectives on Experience. Boston Consulting Group Inc., Boston
4. Brockhoff K (1987) Budgetierungsstrategien füur Forschung und Entwicklung. Z Betriebswirtschaft 75:846–869
5. Brynjolfsson E, Kemerer CF (1996) Network Externalities in Microcomputer Software: An Econometric Analysis of the Spreadsheet Market. Manag Sci 42(12):1627–1647
6. Church J, Gandal N (1993) Complementary network externalities and technological adoption. Int J Ind Organ 11:239–260
7. Farrell J, Saloner G (1986) Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation. Am Econ Review 76(5):940–955
8. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
9. Jeuland AP, Dolan RJ (1982) An Aspect of New Product Planning: Dynamic Pricing. In: Zoltners AA (ed) TIMS Studies in the Management Sciences 18. North Holland, Amsterdam, pp 1–21
10. Katz ML, Shapiro C (1985) Network Externalities, Competition, and Compatibility. Am Econ Review 75(3):424–440
11. Kotler P (1994) Marketing Management. Prentice-Hall, Englewood Cliffs
12. Leibenstein H (1950) Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. Q J Econ 64(2):183–207
13. Mahajan V, Muller E (1979) Innovation Diffusion and New Product Growth Models in Marketing. J Mark 43:55–68
14. Maier FH (1998) New Product Diffusion Models in Innovation Management – A System Dynamics Perspective. Syst Dyn Review 14:285–308
15. Milling P (1996) Modeling Innovation Processes for Decision Support and Management Simulation. Syst Dyn Review 12(3):221–234
16. Milling P, Maier F (1996) Invention, Innovation und Diffusion. Duncker & Humboldt, Berlin
17. Milling P, Maier F (2004) R&D, Technological Innovations and Diffusion. In: Barlas Y (ed) System Dynamics: Systemic Feedback Modeling for Policy Analysis. In: Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO. EOLSS-Publishers, Oxford, p 39; http://www.eolss.net
18. Schmalen H (1989) Das Bass-Modell zur Diffusionsforschung – Darstellung, Kritik und Modifikation. Schmalenbachs Z betriebswirtschaftliche Forsch (zfbf) 41:210–225
19. Senge PM (1994) Microworlds and Learning Laboratories. In: Senge PM et al (eds) The Fifth Discipline Fieldbook. Doubleday, New York, pp 529–531
20. Sterman JD (1992) Teaching Takes Off – Flight Simulators for Management Education. In: OR/MS Today, October 1992. Lionheart, Marietta, pp 40–44
21. Sterman JD (2000) Business Dynamics. Irwin McGraw-Hill, Boston
22. Thun JH, Größler A, Milling P (2000) The Diffusion of Goods Considering Network Externalities – A System Dynamics-Based Approach. In: Davidsen P, Ford DN, Mashayekhi AN (eds) Sustainability in the Third Millennium. Systems Dynamics Society, Albany, pp 204.1–204.14
23. Xie J, Sirbu M (1995) Price Competition and Compatibility in the Presence of Positive Demand Externalities. Manag Sci 41(5):909–926

Books and Reviews

Abernathy JW, Utterback JM (1988) Patterns of Industrial Innovation. In: Burgelman RA, Maidique MA (eds) Strategic Management of Technology and Innovation. Homewood, Irwin, pp 141–148

Bailey NTJ (1957) The Mathematical Theory of Epidemics. Griffin, London

Bass FM (1980) The Relationship between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations. J Bus 53:50–67

Bower JL, Hout TM (1988) Fast-Cycle Capability for Competitive Power. Harvard Bus Review 66(6):110–118

Bye P, Chanaron J (1995) Technology Trajectories and Strategies. Int J Technol Manag 10(1):45–66

Clarke DG, Dolan RJ (1984) A Simulation Analysis of Alternative Pricing Strategies for Dynamic Environments. J Bus 57:179–200

Dumaine B (1989) How Managers can Succeed Through Speed. Fortune 4 February 13:30–35

Easingwood C, Mahajan V, Muller E (1983) A Non-Uniform Influence Innovation Diffusion Model of New Product Acceptance. Mark Sci 2:273–295

Fisher JC, Pry RH (1971) A Simple Substitution Model of Technological Change. Technol Forecast Soc Chang 3:75–88

Ford DN, Sterman JD (1998) Dynamic Modeling of Product Development Processes. Syst Dyn Review 14:31–68

Forrester JW (1968) Industrial Dynamics – After the First Decade. Manag Sci 14:389–415

Forrester JW (1981) Innovation and Economic Change. Futures 13(4):323–331

Georgescu-Roegen N (1971) The Entropy Law and the Economic Process. Harvard University Press, Cambridge

Graham AK, Senge PM (1980) A Long Wave Hypothesis of Innovation. Technol Forecast Soc Chang 17:283–311

Homer JB (1983) A Dynamic Model for Analyzing the Emergence of New Medical Technologies. Ph.D. Thesis. MIT Sloan School of Management

Homer JB (1987) A Diffusion Model with Application to Evolving Medical Technologies. Technol Forecast Soc Chang 31(3):197–218

Kern W, Schröder HH (1977) Forschung und Entwicklung in der Unternehmung. Rowohlt Taschenbuch Verlag, Reinbek

Linstone HA, Sahal D (eds) (1976) Technological Substitution. Forecast Techniques Appl. Elsevier, New York

Maier FH (1992) R&D Strategies and the Diffusion of Innovations. In: Vennix JAM (ed) Proceedings of the 1992 International System Dynamics Conference. System Dynamics Society, Utrecht, pp 395–404

Maier FH (1995) Die Integration wissens- und modellbasierter Konzepte zur Entscheidungsunterstützung im Innovationsmanagement. Duncker & Humboldt, Berlin

Maier FH (1995) Innovation Diffusion Models for Decision Support in Strategic Management. In: Shimada T, Saeed K (eds) System Dynamics '95. vol II. System Dynamics Society, Tokyo, pp 656–665

Maier FH (1996) Substitution among Successive Product Generations – An Almost Neglected Problem in Innovation Diffusion Models. In: Richardson GP, Sterman JD (eds) System Dynamics '96. System Dynamics Society, Boston, pp 345–348

Mansfield EJ, Rapoport J, Schnee S, Wagner S, Hamburger M (1981) Research and Innovation in the Modern Corporation: Conclusions. In: Rothberg RR (ed) Corporate Strategy and Product Innovation. Norton, New York, pp 416–427

Meieran ES (1996) Kostensenkung in der Chip-Fertigung. Siemens Z Special FuE Frühjahr:6–10

Milling P (1986) Diffusionstheorie und Innovationsmanagement. In: Zahn E (ed) Technologie- und Innovationsmanagement. Duncker & Humboldt, Berlin, pp 49–70

Milling PM (1986) Decision Support for Marketing New Products. In: Aracil J, Machuca JAD, Karsky M (eds) System Dynamics: On the Move. The System Dynamics Society, Sevilla, Spain, pp 787–793

Milling PM (1987) Manufacturing's Role in Innovation Diffusion and Technological Innovation. In: Proceedings of the 1987 International Conference of the System Dynamics Society. The System Dynamics Society, Shanghai, China, pp 372–382

Milling PM (1989) Production Policies for High Technology Firms. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Society for Computer Simulation International, Edinburgh, pp 233–238

Milling PM (1991) An Integrative View of R&D and Innovation Processes. In: Mosekilde E (ed) Modelling and Simulation. Simulation Councils, San Diego, pp 509–514

Milling PM (1991) Quality Management in a Dynamic Environment. In: Geyer F (ed) The Cybernetics of Complex Systems – Self-organization, Evolution, and Social Change. InterSystems Publications, Salinas, pp 125–136

Milling PM, Maier FH (1993) Dynamic Consequences of Pricing Strategies for Research and Development and the Diffusion of Innovations. In: Zepeda E, Machuca JAD (eds) The Role of Strategic Modeling in International Competitiveness – System Dynamics '93. The System Dynamics Society, Cancun, Mexico, pp 358–367

Milling PM, Maier FH (1993) The Impact of Pricing Strategies on Innovation Diffusion and R&D Performance. Syst Dyn Int J Policy Model 6:27–35

Norton JA, Bass FM (1987) A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Manag Sci 33:1069–1086

Norton JA, Bass FM (1992) Evolution of Technological Generations: The Law of Capture. Sloan Manag Review 33:66–77

Paich M, Sterman JD (1993) Boom, Bust, and Failure to Learn in Experimental Markets. Manag Sci 39:1439–1458

Pearl R (1924) Studies in Human Biology. Williams and Wilkins, Baltimore

Robinson B, Lakhani C (1975) Dynamic Price Models for New Product Planning. Manag Sci 21:1113–1122

Roberts EB (1964) The Dynamics of Research and Development. Harper and Row Publishers, New York Evanston London

Roberts EB (1978) Research and Development Policy Making, In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 283–292

Roberts EB (1978) A Simple Model of R&D Project Dynamics. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 293–314

Rogers EM (1983) Diffusion of Innovation, 3rd edn. The Free Press, New York

Schumpeter JA (1961) Konjunkturzyklen – Eine theoretisch, historische und statistische Analyse des kapitalistischen Prozesses, Erster Band. Vandenhoeck & Ruprecht, Göttingen

Steele LW (1989) Managing Technology. McGraw-Hill, New York

Sterman JD (1989) Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Environment. Manag Sci 35:321–339

Sterman JD (1994) Learning in and about Complex Systems. Syst Dyn Review 10:291–330

Weil HB, Bergan TB, Roberts EB (1978) The Dynamics of R&D Strategy. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 325–340

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation

KHALID SAEED
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary

Absentee owners Parties not present on land and capital resources owned by them.

Artisan owners Parties using own labor together with land and capital resources owned by them to produce goods and services.

Behavioral relations Causal factors influencing a decision.

Capital intensive A process or industry that requires large sums of financial resources to produce a particular good.

Capital Machinery equipment, cash and material inputs employed for the production of goods and services.

Capitalist sector A subeconomy in which all resources are privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Capitalist system An economic system in which all resources are in theory privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Commercial Pertaining to buying and selling with intent to make profit.

Controlling feedback A circular information path that counters change.

Corporate pertaining to a profit maximizing firm.

Economic dualism Side-by-side existence of multiple subeconomies.

Economic sector A collection of production units with common characteristics.

Entrepreneurship Ability to take risk to start a new business.

Feedback loops Circular information paths created when decisions change information that affects future decisions.

Financial market A mechanism that allows people to easily buy and sell commodities, financial instruments and other fungible items of value at low transaction costs and at prices that reflect efficient markets.

Household income Income accrued to a household from wages, profits and rents received by all its members.

Institutionalist economic models Models attributing performance of economies to institutional relationships and advocating selective government intervention to change the behavior that creates dysfunctions.

Iron law of wages David Ricardo's most well-known argument about wages"naturally" tending towards a minimum level corresponding to the subsistence needs of the workers.

Keynesian A belief that the total spending in the economy is influenced by a host of economic decisions – both public and private.

Labor intensive A process or industry with significant labor costs.

Labor productivity Output per worker or worker-hour.

Labor Economically active persons in an economy.

Marginal factor cost The incremental costs incurred by employing one additional unit of input.

Marginal revenue product The additional income generated by using one more unit of input.

Market economy An economy which relies primarily on interactions between buyers and sellers to allocate resources.

Marxist economic theory A theory highlighting exploitive mechanisms in an economic system and advocating central governance.

Marxist system A centrally run economic system emphasizing in theory the Marxist axiom "from each according to ability to each according to need".

Model An abstract representation of relationships in a real system.

Neoclassical economic theory A theory highlighting constructive market forces in an economic system and advocating consumer sovereignty and a price system as invisible sources of governance.

Non-linear A system whose behavior can't be expressed as a sum of the behaviors of its parts.

Opportunity cost Real value of resources used in the most desirable alternative, or the amount of one commodity foregone when more of another is consumed.

Ordinary differential equation A relation that contains functions of only one independent variable, and one or more of its derivatives with respect to that variable.

Output elasticity Change in output caused by addition of one unit of a production factor.

Perfect market A hypothetical economic system that has a large number of buyers and sellers – all price takers trading a homogeneous product – with complete information on the prices being asked and offered in other parts of the market; and with perfect freedom of entry to and exit from the market.

Political economy Interaction of political and economic institutions, and the political environment.

Production factor A resource input such as land, labor, or capital contributing to production of output.

Productivity The amount of output created (in terms of goods produced or services rendered) per unit input used.

Purchasing power parity The value of a fixed basket of goods and services based on the ratio of a countries' price levels relative to a country of reference.

Revisionist economic models Models recognizing both constructive and exploitive forces and advocating government intervention against exploitation.

Sector A collection of production units with common characteristics.

Self-employment Work for a self-owned production unit without a defined wage.

System dynamics A methodology for studying and managing complex feedback systems, such as one finds in business and other social systems.

Subeconomy A collection of production units and households with common characteristics.

Theories of value How people positively and negatively value things and concepts, the reasons they use in making their evaluations, and the scope of applications of legitimate evaluations across the social world.

Unearned income Income received as rents.

Wage employment Work for a defined wage.

Definition of the Subject

Poverty is perhaps the most widely written about subject in economic development, although there is little agreement over its causes and how to alleviate it. The undisputed facts about poverty are that it is pervasive, growing and that the gap between the rich and the poor is widening.

It is widely believed that the governments – irrespective of their ideological inclinations – have the responsibility to intervene to help the poor. Poverty alleviation is also the key mandate of International Bank for Reconstruction and Development (World Bank) and the many civil society organizations. World Bank places poverty line at purchasing power parity of $1 per day, which has improved a bit in terms of percentage below over the past three decades, except in Africa, but remains large in terms of head count. This threshold is however unrealistic since it targets absolutely basket cases. A poverty line at purchasing power parity of $3 per day, which is close to average purchasing power per capita in the poor countries shows that both poverty head count and gap between rich and poor have been expanding across board. World Bank web site at http://iresearch.worldbank.org/PovcalNet/jsp/index.jsp allows making such computations for selected countries, regions, years and poverty lines.

Neoclassical economic theory does not explicitly address the process of income distribution among households, although it often views income distribution as shares of profits and wages. In most economic surveys and censuses, however, income distribution is invariably measured in terms of shares of various percentages of the households. The fact that more than 80% of the income is claimed by fewer than 20% of the households who also own most of the capital resources in almost all countries of the world, the theory and the measurement have some common ground. Neoclassical theory has, however, shed little light on the process of concentration of wealth and how can this dysfunction be alleviated.

System dynamics, although rarely used for the design of public policy for addressing poverty, allows us to construct and experiment with models of social systems to understand their internal trends and test policy combinations for changing them. In this paper I have used system dynamics modeling to understand the process of concentration of wealth and re-evaluate the on-going poverty alleviation effort.

The model, which subsumes resource allocation, production and entitlements, explains the many manifestations of income distribution in a market economy. It generates multiple patterns of asset ownership, wage and employment assumed in neo-classical, Marxist and revisionist perspectives on economic growth while it allows ownership to change through the normal course of buying and selling transactions based on rational though, information-bound criteria. Privately owned resources can be employed through hiring wage-labor, rented out or used for self-employment. In addition to the labor market conditions, the wage rate depends also on the opportunity cost of accepting wage employment as workers may be either self-employed or wage-employed. Since this opportunity cost varies with the capital resources owned by the workers, which may support self-employment, the wage rate is strongly affected by the distribution of ownership. Thus, ownership can become concentrated or widely distributed depending on legal and social norms governing transactions in the economy, which the model replicates. Extended experimentation with this model serves as a basis to identify critical policy instruments that make best use of the system potential for resource constrained growth and poverty alleviation through widening participation in the market and improving income distribution.

Introduction

The opening up of the major centrally planned economies of the world has brought to the fore problems concerning the psychological deprivation, inefficiencies of resource allocation and production, and the lack of dynamism experienced in the working of central planning in a socialist system. The accompanying enthusiasm for free market in a capitalist system has, however, hidden many of the dysfunctional aspects of this alternative. It should be recognized that both systems emerged from time-specific and geography-specific empirical evidence. Since their underlying models treat as given specific economic patterns, the institutional roles and the legal norms associated with each system have inherent weaknesses, which create dysfunctions when implemented in different environmental contexts [43,64]. Thus, neither model may furnish an adequate basis for the design of policies for sustainable economic development and poverty alleviation. A search is, therefore, necessary for an organizational framework that might explain the internal trends inherent in each model as special modes of a complex system subsuming the variety of behavioral patterns recognized by specific models before an effective policy for change can be conceived [52].

Using as an experimental apparatus a formal model of the decision structure affecting wage determination, saving and investment behavior, and the disbursement of income, presented earlier in [53], this paper seeks to identify the fundamental economic relations for creating a dynamic and sustainable market system that may also increase opportunities for the poor, whose market entry is often limited by their financial ability and social position [58], to participate in the economy and be entitled to the value it creates. System dynamics provides the technical framework to integrate the various behavioral relations in the system [13,63].

Notwithstanding the many objections to the abstract models of orthodox economics, which are difficult to identify in the real world [28,46], the model of this paper draws on neo-classical economics to construct a basic structure for growth and market clearing. This structure is, however, progressively modified by relaxing its simplifying assumptions about aggregation of sub-economies, wage determination, ownership, income disbursement, saving and investment behavior, financial markets, and technological differentiation between sub-economies to realize the many growth and income distribution patterns addressed in a variety of economic growth models.

The modified model I finally create represents a real world imperfect market in which expectations formed under bounded rational conditions govern the decisions of the economic actors [59], as recognized in the pioneering works of Kaldor (1969) [24], Kalecki (1965) [25], Wientraub (1956) [66], and Joan Robinson (1978) [45]. The model also subsumes the concept of economic dualism first recognized by Boeke (1947) [7] and developed further by Lewis (1958) [29], Sen (1966) [57], Bardhan (1973) [4] and others to represent the multiple sub-economies that co-exist especially in the developing countries. Such a model is more identifiable with respect to the real world as compared with the time and geography specific concepts propounded by the various, often controversial, theories of economic growth.

Simulation experiments with the model explore entry points into the economic system for creating an egalitarian wage and income distribution pattern through indirect policy instruments. Also explored are the functions of entrepreneurship and innovation and the mechanisms that may increase the energy of those processes toward facilitating economic growth and alleviating poverty.

The Alternative Economic Models and Their Limitations

The economic models used as the bases for designing development policies over the past several decades have ascended largely from time-specific and geography-specific experiences rather than from a careful study of the variety of behavioral patterns occurring over various time periods and across several geographic locations. Among these, the socialist and the capitalist models are most at odds. They differ in their assumptions about ownership and income distribution patterns, the basis for wage determination, the influence of technology on income growth and the functions of entrepreneurship and innovation [21,55].

The neo-classical economic theory, which is the basis for the capitalist model, is silent on the ownership of capital resources, often assuming it in default to be widely distributed [5]. Thus, the labor-wage rate may bear little relationship to the income of households, who are also recipients of profits. It is assumed that private control of productive resources is a means for market entry, which creates unlimited potential for economic growth, although private investment is not often seen to be subject to self-finance due to the assumption that financial markets are perfect. The neo-classical economic theory also postulates that short-run labor-wage rates depend on labor market conditions, while in the long run, they are determined by the marginal revenue product of labor. Neo-classical models of economic growth, however, often make the simplifying assumption that equilibrium continues to prevail in both factor and product markets over the course of growth. Thus, only minor fluctuations may occur in wages, profits and prices in the short run, and these can be ignored.

The belief in the existence of such equilibrium is further strengthened by the Keynesian argument for the ineffectiveness of the market mechanisms due to the dependence of prices on long-term wage contracts and production plans which may not respond easily to short-run changes of the market. As a result of the belief in this theory of wage determination, technological choices that increase labor productivity are expected to have a positive effect on wage rates and household income, because they increase the marginal revenue product of labor. Entrepreneurship is viewed as important for new entry into economic activity, which is open to all, and innovation is supposed to benefit society through increased productivity. With these assumptions, the capitalist system advocates minimal government intervention in the economy. This model is widely presented in the many texts on economic development. Pioneering texts include Hirshleifer (1976) [22] and Kindelberger and Herrick (1977) [27].

Marxist economic theory, which underpins the socialist model, assumes on the other hand that ownership of capital resources is concentrated in a minority excluding the workers and that the majority of households receive no part of the profits. Thus, wage payments have a strong effect on household income. The Marxist theory views private ownership as a source of exploitation and postulates labor-wage rates determined by the consumption necessary for a worker to support production in a grossly labor surplus economy following Ricardo's iron law of wages [32,39]. The labor-wage rate is, thus, based on the real value of the commodities needed for a worker to subsist, which is more or less fixed, irrespective of the contribution of labor to the production process. In such conditions, technological choices that increase labor productivity may indeed only serve to increase the share of the surplus of product per unit of labor appropriated by the capitalists. In this model, entrepreneurship is viewed as an asocial activity and innovation seen to originate from the need to boost falling returns on capital. Attributing the development of these conditions to market failure, the socialist system assigns control of the economy to the government.

There also exist a number of revisionist models of political economy attempting to explain the nature of interdependence of the multiple sub-economies observed to co-exist in many developing countries in violation of the theoretical premises of the neo-classical model according to which all production factors must eventually move to the most efficient sector. These models often attribute the development of disparities between the various sub-economies to exploitative mechanisms that tend to maintain an upper hand of the stronger influence groups. The revisionist analyses have largely led to making moral appeals for the government policy to target the poor and the disadvantaged in its development effort, which is a stated mandate of the International Bank for Reconstruction and Development (World Bank). Targeting the poor has also been advocated widely by numerous development economists over the past half century. They include such prominent economists as Myrdal (1957) [36], Lipton (1977) [30], Galbraith (1979) [15], and Sen (1999) [58].

Indeed, each economic system can often be endorsed with the help of selected historical evidence, and this has been fully exploited to fuel the traditional debate between the neo-classical and Marxist economic schools. Interesting artifacts of this debate include the normative theories of value suggested by each system to justify the various wage systems, which have little practical significance for development policy [44,62]. This is unfortunate, since contradictions of evidence should clearly indicate the existence of fundamental organizational arrangements in the economic system, which are capable of creating the multiple behavior patterns on which the various economic models are based. Once identified, such arrangements may also serve as entry points for the design of evolutionary changes in an existing pattern. To quote Professor Joan Robinson:

> Each point of view bears the stamp of the period when it was conceived. Marx formed his ideas in the grim poverty of the forties. Marshal saw capitalism blossoming in peace and prosperities in the sixties.

> Keynes had to find an explanation for the morbid condition of 'poverty in the midst of plenty' in the period between the wars. But each has significance for other times, for in so far as each theory is valid, it throws light upon essential characteristics of the system which have always been present in it and still have to be reckoned with. [43]

Following sections of this paper experiment with a system dynamics model of an economic system, widely found in the developing countries and presented earlier in [53], to understand the variety of economic patterns experienced over time and geography under different legal and social norms. Furthermore, exploratory experimentation with this model helps to outline the basic principles of a market system that can sustain growth, create equitable distribution of benefits and facilitate innovation and productivity improvement, all widely deemed necessary for poverty alleviation.

A System Dynamics Model of Resource Allocation, Production and Entitlements

A system dynamics model subsuming the broad decision rules that underlie resource allocation, production, and income disbursement processes of a developing country economic system was proposed in Saeed (1988) [49] and further experimented with in Saeed (1994) [53]. In this model, capital, labor, and land (which may also be assumed as a proxy for natural resources) are used as production factors. Model structure provides for the functioning of two modes of production, commercial, in which resources are employed on the basis of their profitability and which is managed by the capitalist sector of the economy; and self-employed, in which workers not employed in the commercial mode make a living. These two modes of production have been referred to variously in the literature, for example as oligopolistic and peripheral firms [16], formal and informal sectors [29], and modern and traditional subeconomies [12].

It has been assumed in the model that all workers, whether self-employed using their own or rented capital resources or employed as wage-workers by the capitalist sector, are members of a homogeneous socio-economic group with a common interest, which is to maximize consumption. This group is also the sole supplier of labor in the economy since the small number of working capitalists is ignored. On the other hand, the capitalist sector is assumed to maximize profit while it is also the sole wage-employer in the economy [2,3,57].

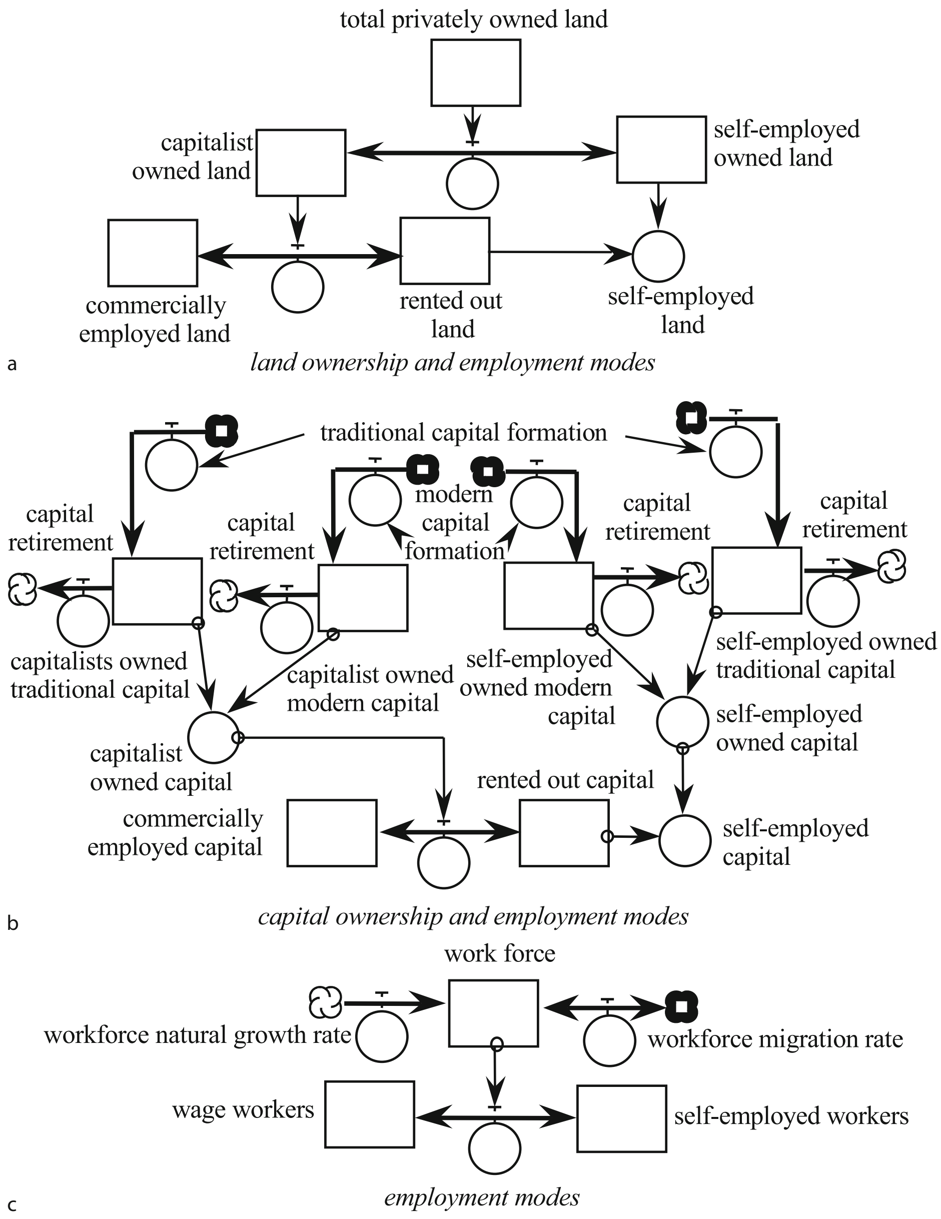
It is assumed that private ownership is protected by law, but land and capital assets can be freely bought, sold and rented by their owners. Each buying and selling transaction between the two sectors must be accompanied by a corresponding transfer of the cash value of the assets determined by the going market prices. The model also permits the appearance of technological differences between the capitalist and the self-employed sectors, when more than one technologies embodied in the type of capital used (named traditional and modern in the model) are available and the two sectors cannot employ the preferred technology with equal ease [30,41], or when the self-employed sector is burdened by excess workers not employed by the commercial sector while it lacks the financial capacity to use its preferred technology.

Figure 1 shows how workers and capital might potentially be retained and employed by the two sectors in the model. Rectangles represent stocks, valve symbols flows and circles intermediate computations following the diagramming convention of system dynamics modeling. The size of each sector is not specified and is determined endogenously by the model, depending on assumptions about the socio-technical environment in which the system functions.

The changes in the quantities of the production factors owned or employed by each sector are governed by the decisions of the producers and the consumers of output and by the suppliers of the production factors acting rationally according to their respective motivations within the bounds of the roles defined for them by the system [59]. The value of production is shared by households on the basis of the quantity of the production factors they contribute and the factor prices they can bargain for [10]. Income share of the workers, less any investment needed to maintain self-employment, divided by the total workforce, determines average consumption per worker, which represents the opportunity cost of accepting wage-employment and this is the basis for negotiating a wage [57,62].

Investment and saving rates in the two sectors are decoupled through a balance of internal savings. The financial markets are segmented by sectors and the investment decisions of a sector are not independent of its liquidity position, given by the unspent balance of its savings. Thus, investment decisions depend on profitability criteria, but are constrained by the balance of internal savings of each sector [33,34]. Figure 2 shows the mechanisms of income disbursement, saving and internal finance incorporated into the model.

The saving propensity of all households is assumed not to be uniform. Since capitalist households receive incomes that are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the


total privately owned land
capitalist owned land
self-employed owned land
commercially employed land
rented out land
self-employed land
a
land ownership and employment modes
traditional capital formation
modern capital formation
capital retirement
capital retirement
capital retirement
capital retirement
capitalists owned traditional capital
capitalist owned modern capital
self-employed owned modern capital
self-employed owned traditional capital
capitalist owned capital
self-employed owned capital
rented out capital
commercially employed capital
self-employed capital
b
capital ownership and employment modes
work force
workforce natural growth rate
workforce migration rate
wage workers
self-employed workers
c
employment modes

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 1
Potential worker and capital distribution between capitalist and self employed sectors

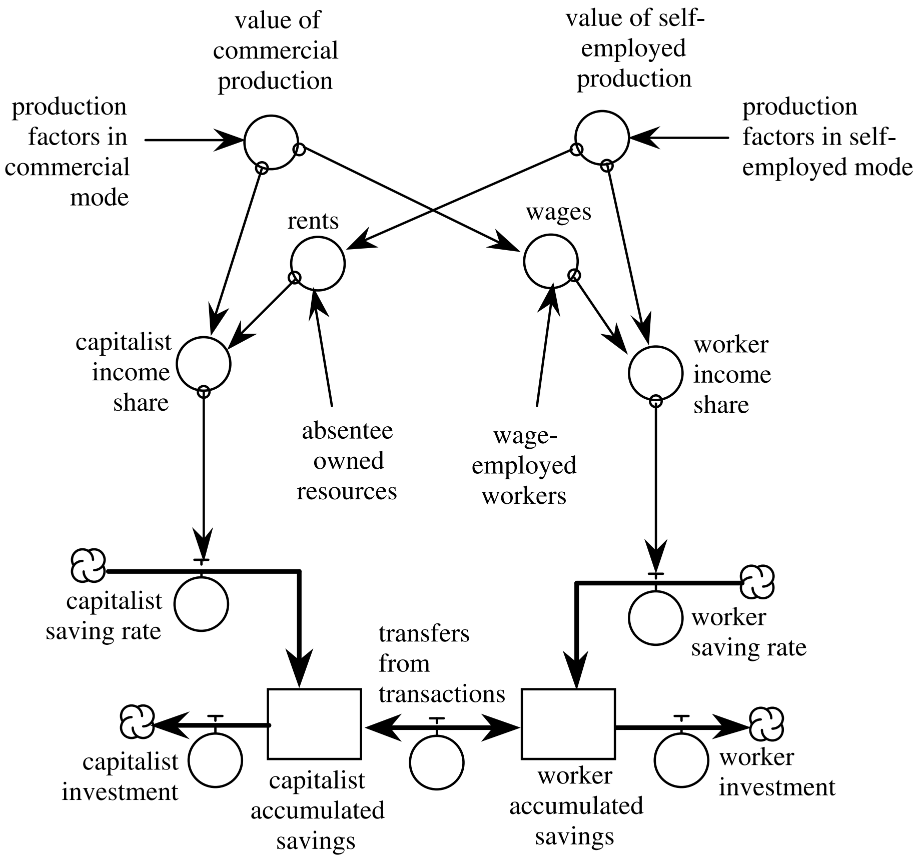
worker households depends on their need to save to support investment for self-employment and on how their absolute level of income compares with their inflexible consumption [23,26,31].

The broad mathematical and behavioral relationships incorporated into the model are given in the Appendix "Model Description". Technical documentation and a machine-readable listing of the model written in DYNAMO code are available from the author on request.

Replicating Income Distribution Patterns Implicit in Models of Alternative Economic Systems

The model is simulated under different assumptions about wages, rents, financial markets and technology and its behavior analyzed in relation to the various theoretical and empirical premises its information relationships represent.

As an arbitrary initial condition, production factors are equally divided between the two sectors and equilib-


value of commercial production
value of self-employed production
production factors in commercial mode
production factors in self-employed mode
rents
wages
capitalist income share
worker income share
absentee owned resources
wage-employed workers
capitalist saving rate
worker saving rate
transfers from transactions
capitalist investment
capitalist accumulated savings
worker accumulated savings
worker investment

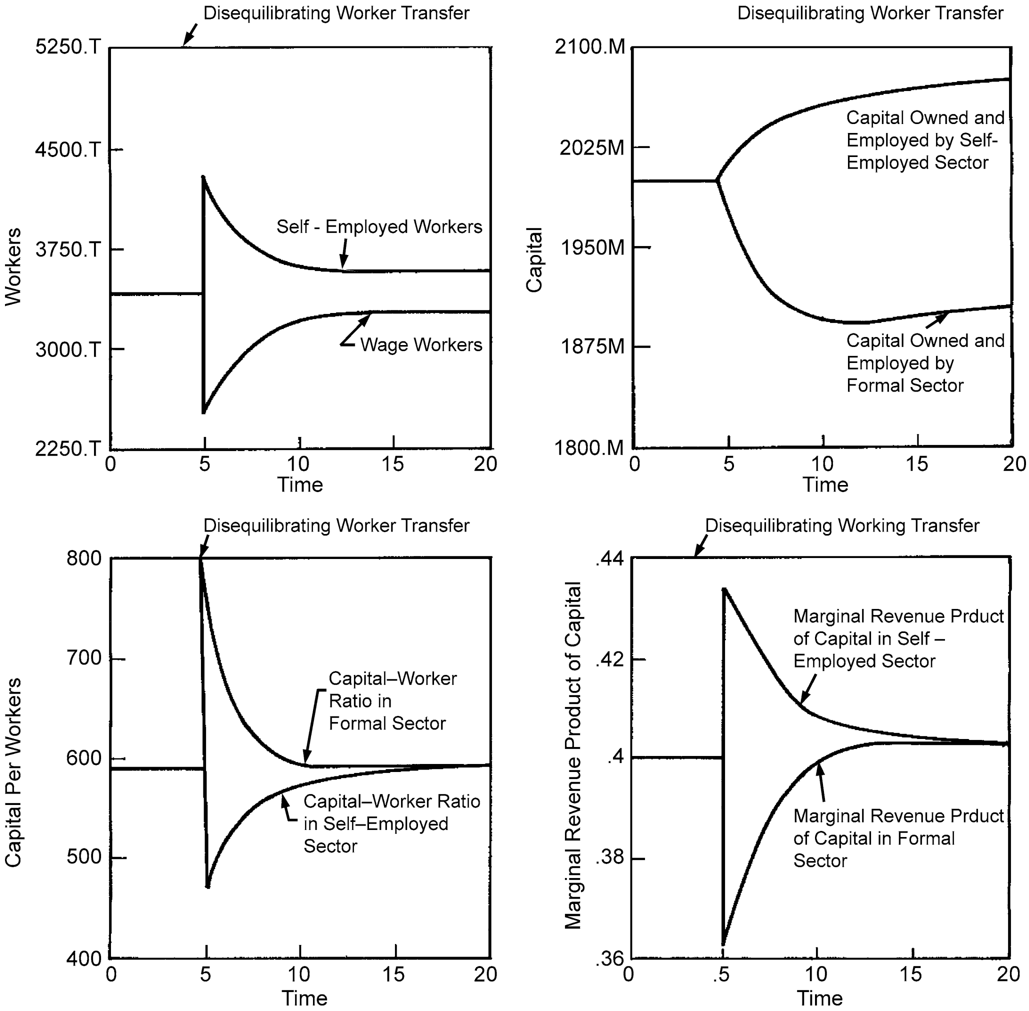
Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 2
Income disbursement process

rium in both product and factor markets is assumed to exist under the conditions of a perfect economic system as described in neo-classical economics. Thus, the marginal revenue products of land and capital are initially assumed to be equal to their respective marginal factor costs determined by an exogenously specified interest rate which represents the general pattern of preferences of the community for current as against future consumption [22]. The marginal revenue product of workers is initially set equal to wage rate. The market is initially assumed to be clear and there is no surplus of supply or demand.

Replicating the Theoretical Neo-classical System

This experiment is aimed at understanding internal trends of a system representing the neo-classical economic theory. To transform the model to represent this system, it is assumed that the production factors employed by each sector are owned by it and no renting practice exists [5]. The wage rate is assumed to be determined by the marginal revenue product of workers and the availability of labor instead of the opportunity cost to the workers of supplying wage-labor. Financial markets are assumed to be perfect and investment decisions of the two sectors are uncoupled from their respective liquidity positions. It is also assumed that the technology of production is the same in the two sectors and, in terms of the model, only traditional capital is available to both of them. The only difference between the two sectors is that the capitalist sector can vary all production factors, including labor to come to an efficient mix, while the self-employed sector may absorb all labor not hired by the capitalist sector, while it can freely adjust other production factors to achieve an efficient mix.

The model thus modified stays in equilibrium when simulated as postulated in neo-classical economic theory. When this equilibrium is disturbed arbitrarily by transferring a fraction of the workers from the capitalist to the self-employed sector, the model tends to restore its equilibrium in a manner also similar to that described by the neo-classical economic theory. This is shown in Fig. 3.



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 3
Recovery from dis-equilibrium in a neo-classical system

The transfer raises the marginal revenue product of workers in the capitalist sector, which immediately proceeds to increase its workforce. The transfer also raises the intensity of workers in the self-employed sector as a result of which the marginal revenue products of land and capital in that sector rise. Hence, it proceeds to acquire more land and capital. These activities continue until the marginal revenue products of the factors and their proportions are the same in the two sectors. Note that while the factor proportions and marginal revenue products of the factors are restored by the model to their original values, the absolute amounts of the various factors are different when new equilibrium is reached. There is, however, no difference in endowments per worker between the capitalist and the self-employed sectors.

Since factor payments are determined purely on the basis of contribution to the production process while the quantities of production factors allocated to each sector depend on economic efficiency, the wages and factor allocations seem to be determined fairly and efficiently, as if by an invisible hand. Ownership in such a situation can either be communal or very widely distributed among households since otherwise the wage bargaining process will not lead to fair wages. Renting of production factors among households is irrelevant since transfer to parties who can efficiently employ them is automatic.

Before anything is said about the empirical validity of the simplifying assumptions made in this model, the historical context of these assumptions must be examined carefully. The simplified model is based on Adam Smith's description of an industrial economy observed at the start of the industrial revolution. This economy was run by artisan-turned capitalists and there were many of these capitalists competing with one another, although, none had the financial muscle to outbid the others except through his/her ability to employ resources efficiently [60].

As far as labor wage rate was concerned, although there were instances of exploitation of workers at a later stage of the industrial revolution, the artisan workers could obtain a wage that was equal to their contribution of labor to the production process, as otherwise they could easily be self-employed since the economy was still quite labor intensive and the tools needed for self-employment may not have cost very much. Also, since ownership of the tools of a trade may have been quite widespread while the contribution of capital resources to the production process was quite small as compared to that of labor, a major part of the income might have accrued to the working households. In such circumstances, the simplifying assumptions of the neo-classical model may appear quite reasonable.

The neo-classical model became irrelevant, however, as the system made progress in the presence of a social organizational framework that legally protected ownership of all types and freely allowed the renting of assets, thus making possible an absentee mode of owning productive resources while technological changes also made the contribution of capital resources to the production process more significant.

Creating Worker Capitalism

It is not only methodologically expedient but also pedagogically interesting to explore what ownership and wage patterns might have emerged if labor-wages were determined through bargaining mechanisms incorporated into the model instead of fair payment equal to the marginal revenue product of workers, while all other assumptions of a perfect market of the experiment of the last section were maintained.

Figure 4 shows a simulation of the model in which wage rate is determined by the average consumption expenditure per worker (as given in Eqs. (1) and (2) of the model described in the Appendix "Model Description") while renting of production factors and financial fragmentation of the households are still not allowed. This change in assumptions disturbs the initial market equilibrium in the model thus activating its internal tendency to seek a new equilibrium. No exogenous disequilibrating changes are needed to generate the dynamic behavior in this simulation and in those discussed hereafter.

As a result of this change, the compensation demanded for working in the capitalist sector becomes much higher than the marginal revenue product of the workers. Thus, wage-workers are laid off and accommodated in the self-employed sector. Consequently, the marginal revenue product of land and capital in the self-employed sector increases and its bids for these resources rise. On the other hand, the decrease in the workforce of the capitalist sector increases its land and capital intensities and hence lowers their marginal revenue products. The falling productivity of these resources increases the opportunity cost of holding them. Since renting is not allowed, the capitalist sector is persuaded to sell the resources to the self-employed who can easily buy them since investment in the model is not subject to internal self-finance.

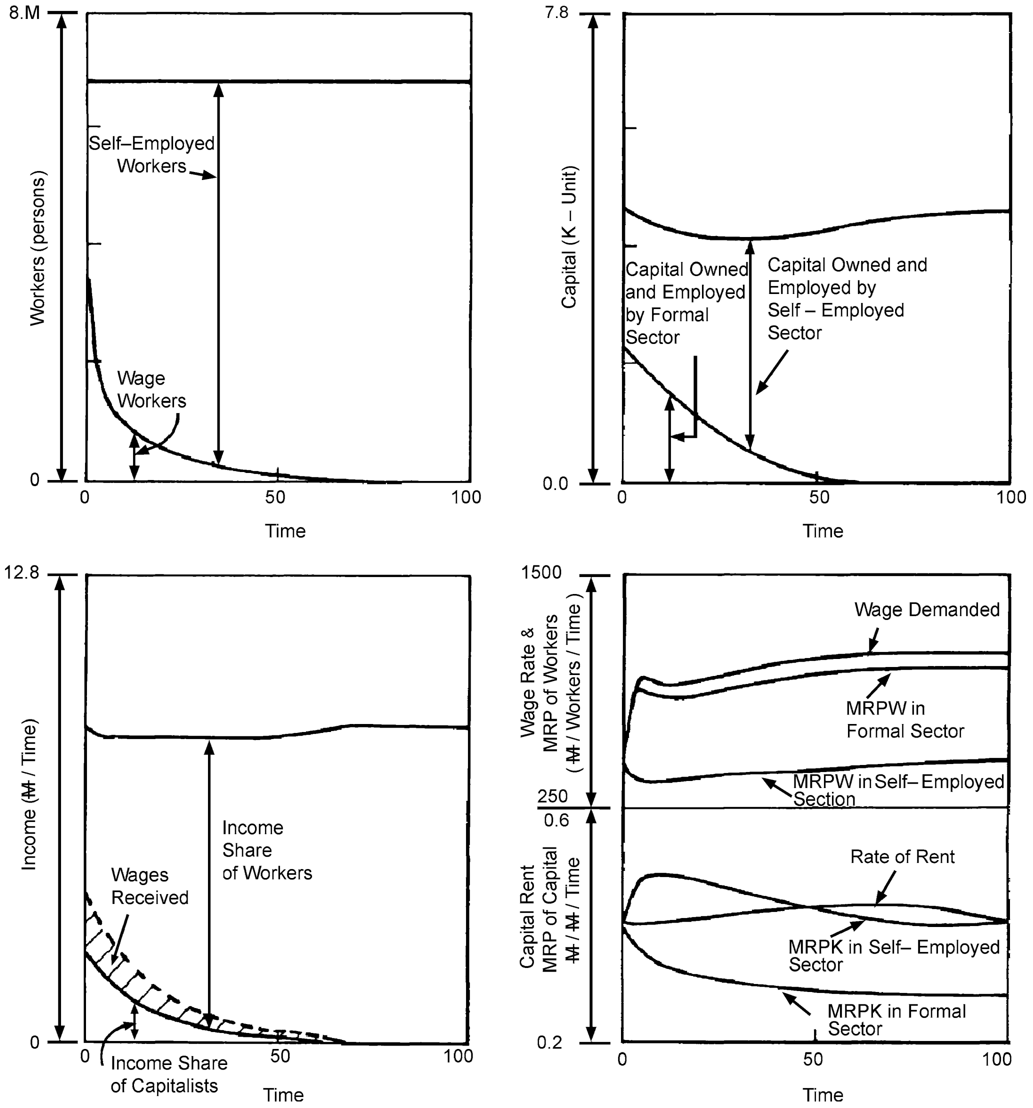
As the self-employed sector increases its land and capital holdings, its production rises. When increases in the production of this sector exceed the wage income lost due to decreasing wage disbursements from the capitalist sector, the net revenue of the workers, and hence their average consumption, rises. The wage rate is thus pushed up further, which necessitates further reductions in wage-workers. These processes spiral into a gradual transfer of all resources to the self-employed sector.

The marginal revenue products of land and labor in the two sectors tend to equilibrate at different values, but the capitalist sector exists only in theory because towards the end of the simulation almost all the resources are owned and managed by the self-employed. Since no part of the income is obtained by absentee owners, and working households may own and manage resources according to the quantity of labor they can supply, the income distribution may appear to be truly egalitarian.

Even though the above simulation is hypothetical, the wage and income distribution pattern shown by it may be experienced when the separation of resources from the households employing them is socially or legally ruled out or the state allocates capital resources and land according to the quantity and quality of labor supplied by a household. Instances of peasant economies having such characteristics have been recorded in history in tribal cultures and, in a somewhat advanced form, in medieval India [35]. Interestingly, such implicit assumptions are also subsumed in the illusive perfect market the neoclassical economic theory is based on.

Appearance of Absentee Ownership

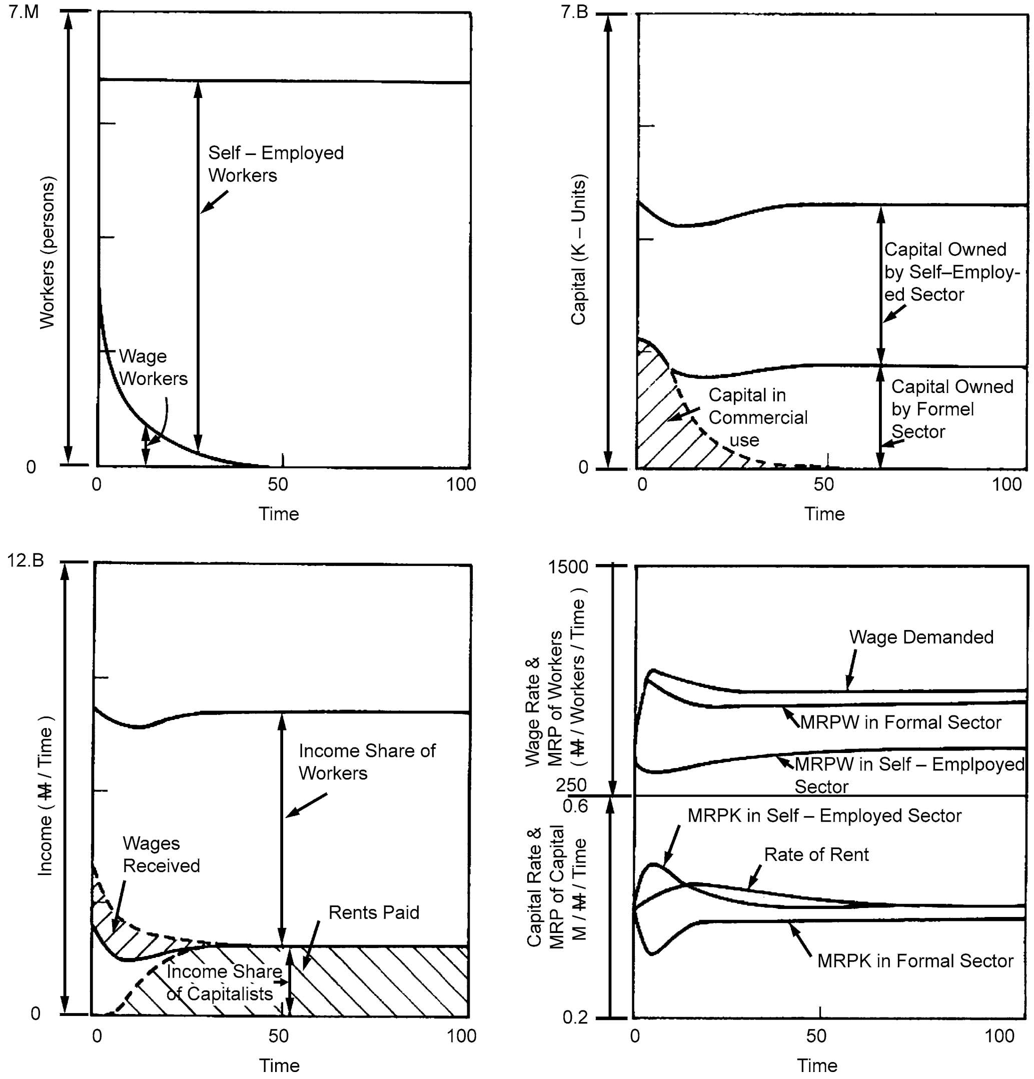
When ownership of resources is legally protected, whether they are productively employed or owned in absentia, many renting and leasing arrangements may appear which



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 4
The develop of worker capitalism when wages depend on bargaining position of workers

may allow a household to own resources without having to employ them for production [47]. This is borne out in the simulation of Fig. 5, in which resources are divided by the capitalist sector between commercial production and renting activities depending on the rates of return in each. Rents depend on long-term averages of the marginal revenue products of the respective factors and on the demand for renting as compared with the supply of rentable assets. In the new equilibrium reached by the model, the commercial mode of production and wage-employment gradually disappear but a substantial part of the resources continues to be owned by the capitalist sector, which rents these out to the self-employed sector.

Such a pattern develops because of the combined effect of wage and tenure assumptions incorporated into the model. When workers are laid off by the capitalist sector in response to a high wage rate, the marginal revenue products of land and capital for commercially employing these resources in this sector fall. However, as the laid-off workers are accommodated in the self-employed sector, the marginal revenue products of land and capital, and hence their demand in this sector, rise. Therefore, rents are bid up and the capitalist sector is able to get enough return



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 5
The appearance of absentee ownership when renting is also allowed

from renting land and capital to justify maintaining its investment in these.

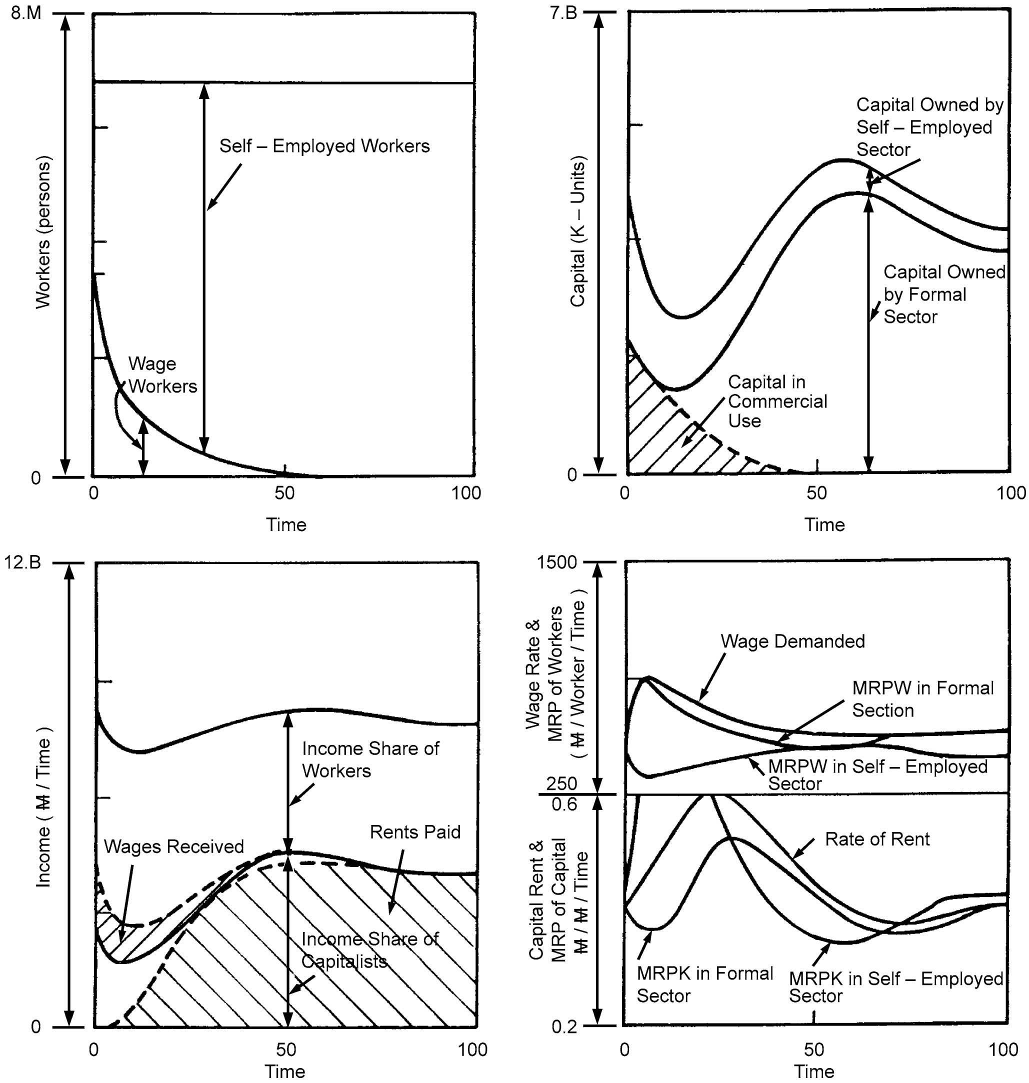
Again, the marginal revenue products of the production factors in the commercial mode of production are only hypothetical as that mode is not practiced towards the end of the simulation. The renting mechanism allows the self-employed sector to adjust its factor proportions quickly when it is faced with the accommodation of a large number of workers. When the economy reaches equilibrium, the marginal rates of return of the production factors in the self-employed sector are the same as those at the beginning of the simulation. But, the wage demanded equilibrates at a level lower than that for the exclusively self-employed economy described in the simulation of Figure 4, because a part of the income of the economy is now being obtained by the absentee owners of the capitalist sector in the form of rent.

Note that, although the total income of the economy falls a little during the transition, it rises back to the original level towards the end equilibrium since the technology is uniform, irrespective of the mode of production. Also note that the end equilibrium distribution of income depends on initial distribution of factors when modifying assumptions are introduced, and on the volume of transfers occurring over the course of transition. Thus, an unlimited number of income and ownership distribution patterns would be possible depending on initial conditions and the parameters of the model representing the speeds of adjustment of its variables. The common characteristics of these patterns, however, are the presence of absentee ownership, the absence of a commercial mode of production, and a shadow wage that is less than an exclusively self-employed system.

Separation of Ownership from Workers and the Creation of a Marxist System

The ownership of resources becomes separated from the workers and concentrated in the capitalist sector in the model, irrespective of the initial conditions of resource distribution, when the assumption about the existence of a perfect financial market is also relaxed.

Figure 6 shows the ownership and wage pattern which develops when acquisition of resources by the capitalist and self-employed sectors is made dependent, in addition to their profitability, on the ability to self-finance their purchase. Recall also that the ability to self-finance depends on the unspent balance of savings, and the saving rate of the self-employed sector is sensitive both to the utility of saving in this sector to support investment for self-employment and to the rent burden of this sector compared with the factor contribution to its income from land and capi-



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 6
Separation of ownership from workers as postulated by Marx system when investment must also be internally financed

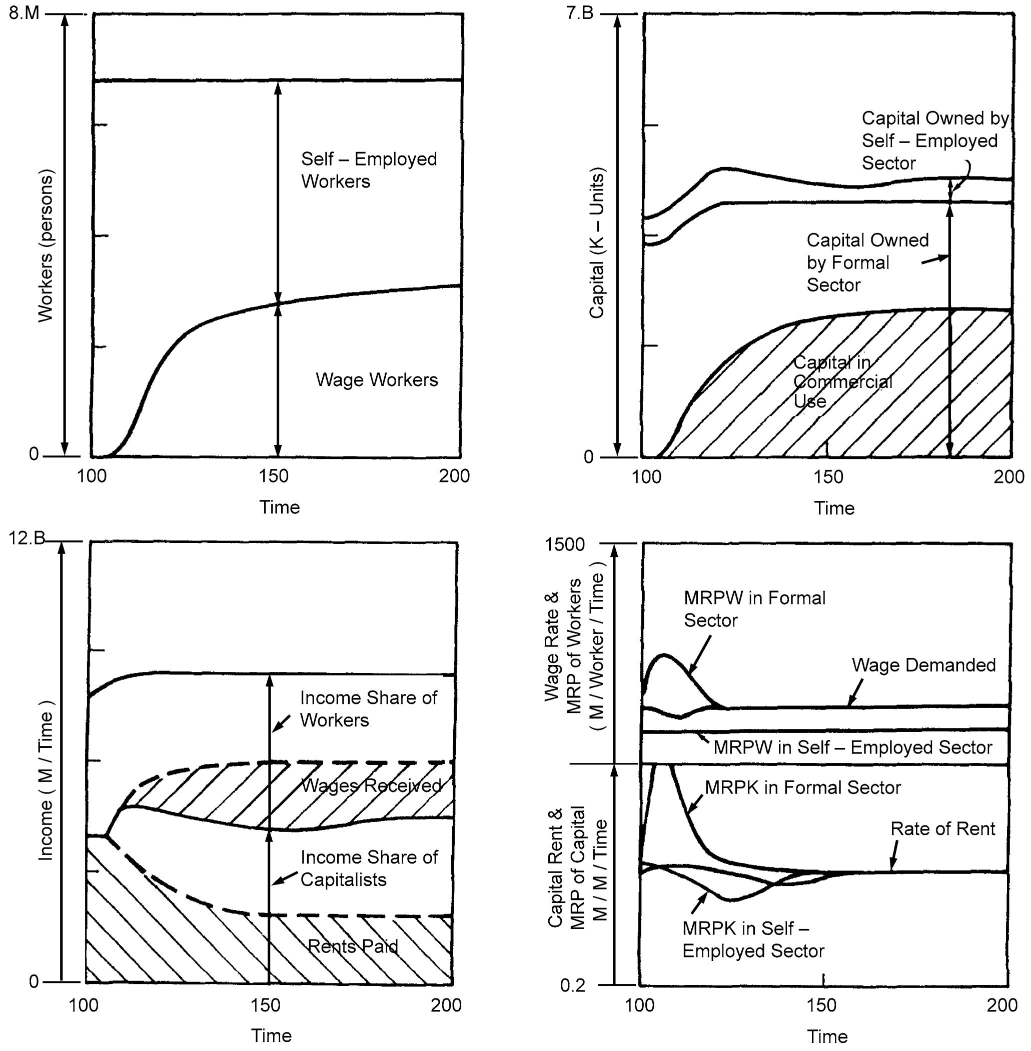
tal. The saving rate of the capitalist sector is assumed to be constant.

Such a pattern develops because of an internal goal of the system to employ resources in the most efficient way while the ownership of these resources can only be with the households who have adequate financial ability, which is also not independent of ownership.

Creation of a Dualist System

A dualist system is characterized by the side-by-side existence of both commercial and self-employed modes of production. The former appears to be economically efficient and is often also capital-intensive. The latter is seen to be economically inefficient and is also invariably labor-intensive. The side-by-side existence of these two modes of production in many developing countries has often puzzled observers, since according to the neo-classical economic theory, any inefficient production mode must be displaced by the efficient one.

A stable commercially run capital-intensive production sector existing together with a self-employed labor-intensive sector develops in the model if a technological differentiation is created between the capitalist and self-employed sectors. This is shown in the simulation in Fig. 7, in which an exogenous supply of modern capital is made



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 7
Creation of dualist system when technological differentiation develops between the capitalist and self-employed sectors

available after end equilibrium of the simulation in Fig. 6 is reached.

Capital differentiation between the two sectors appears since the scale of the self-employed producers does not allow them to adopt modern technologies requiring indivisible capital inputs. The capitalist sector starts meeting its additional and replacement capital needs by acquiring a mixture of modern and traditional capital while the self-employed sector can use only traditional capital. However, the capital demand of the capitalist sector is met by modern capital as much as the fixed supply permits. The balance of its demand is met by acquiring traditional capital.

The output elasticity of modern capital is assumed to be higher than that of the traditional capital while the use of the former also allows an autonomous increase in output. The output elasticity of land is assumed to remain constant. The assumption of uniform returns to scale is maintained. Thus, the output elasticity of workers decreases when modern capital is introduced. These assumptions serve to represent the high productivity and labor-saving characteristics of the modern capital.

As its capital becomes gradually more modern and potentially more productive, the capitalist sector is able to employ its productive resources with advantage in the commercial mode of production, instead of renting these out, and to employ wage-workers at the going wage rate. The increased productivity and income derived from this make it both economically and financially viable for the capitalist sector to invest more. Thus, its share of resources, when a new equilibrium is reached, is further increased.

Since the output elasticity of workers falls with the increase in the fraction of modern capital, the marginal revenue product of workers in the commercial mode may not rise much with the increase in its output. At the same time, since resources are being transferred away by the capitalist sector from renting to commercial employment, the labor intensity and the demand for renting rises in the self-employed sector. Hence rents are bid up and it again becomes profitable for the capitalist sector to allocate resources to renting. The amount of resources rented out, however, will depend on the degree of technological differentiation that may be created between the two sectors.

The wage rate reaches equilibrium at a lower level and the rents at higher levels than without technological differentiation. Rents, however, equal marginal revenue products of land and capital, which rise in the capitalist sector because of employing superior technology and in the self-employed sector due to increased labor intensity.

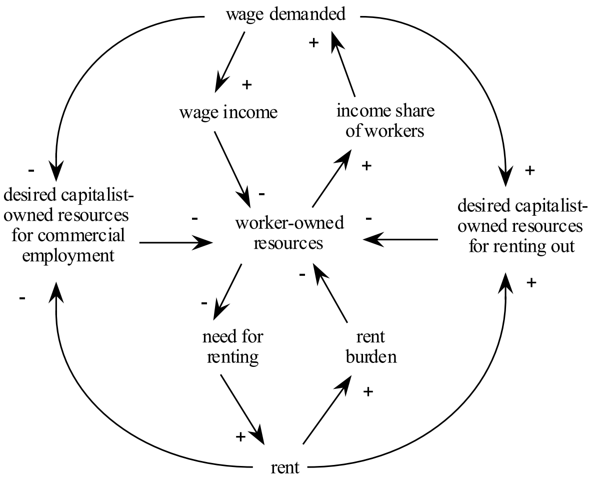
Interestingly, dualist patterns appeared in the developing countries, both in the agricultural and industrial sectors, only after modern capital inputs became available in limited quantities. Labor-intensive peasant agriculture and small-scale industry and services carried out by the self-employed came to exist side-by-side with the commercially run farms and large-scale industry employing wage labor and modern technologies. However, worker income, both in wage-employment and self-employment, remained low [19].

Feedback Loops Underlying Wage and Income Patterns

The internal goal of a dynamic system represented by a set of non-linear ordinary differential equations is created by the circular information paths or feedback loops which are formed by the causal relations between variables implicit in the model structure. These causal relations exist in the state space independently of time (unless time also represents a state of the system). The existence of such feedback loops is widely recognized in engineering and they are often graphically represented in the so-called block and signal flow diagrams [17,40,65].

While many feedback loops may be implicit in the differential equations describing the structure of a system, only a few of these would actively control the system behavior at any time. The nonlinearities existing in the relationships between the state variables determine which of the feedback loops would actively control the system behavior. A change may occur in the internal goals of a system if its existing controlling feedback loops become inactive while simultaneously other feedback loops present in its structure become active. Such a shift in the controlling feedback loops of a system is sometimes called a structural change in the social sciences and it can result both from the dynamic changes occurring over time in the states of the system and from policy intervention. The realization of a specific wage and income distribution pattern depends not on assumptions about initial conditions but on legal and social norms concerning ownership, renting, financing of investment and the state of technology, determining which feedback loops would be dominant [14,40].

Figure 8 describes the feedback loops, formed by the causal relations implicit in the model structure that appear to polarize income distribution by separating asset ownership from working households and creating a low wage rate, as shown in Fig. 6. An arrow connecting two variables indicates the direction of the causality while a positive or a negative sign shows the slope of the function relating cause to effect. For clarity, only key variables located along each feedback path are shown.


wage demanded
wage income
income share of workers
desired capitalist-owned resources for commercial employment
worker-owned resources
desired capitalist-owned resources for renting out
need for renting
rent burden
rent
+
-

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 8
Feedback loops creating dysfunctional income distribution trends in the capitalis system

When productive resources can potentially be engaged in commercial or self-employed modes by owners and renters, any autonomous increase in the wage rate would not only decrease the desired capitalist owned resources for commercial employment, it would also concomitantly decrease the utility of investing in resources for self-employment. Thus, while the ownership of resources freed from commercial employment is not transferred to the self-employed sector, the surplus labor released by the commercial sector has to be absorbed in self-employment. As a result, worker income is depressed while the demand for renting rises. Thus, it not only continues to be profitable for the capitalist sector to hold its investments in land and capital, it also gives this sector a financial edge over the self-employed sector, whose savings continue to decline as its rent burden rises. These actions spiral into an expansion of ownership of resources by the capitalist sector even though the commercial mode of production is eliminated due to the high cost of wage labor. This also precipitates a very low wage rate when equilibrium is reached since a low claim to income of the economy creates low opportunity costs for the self-employed workers for accepting wage-employment.

Ironically, the fine distinction between the corporate, artisan and absentee types of ownership is not recognized in the political systems based on the competing neoclassical and Marxist economic paradigms. The former protects all types of ownership; the latter prohibits all. None creates a feasible environment in which a functional form of ownership may help to capture the entrepreneurial energy of the enterprise.

Possibilities for Poverty Alleviation

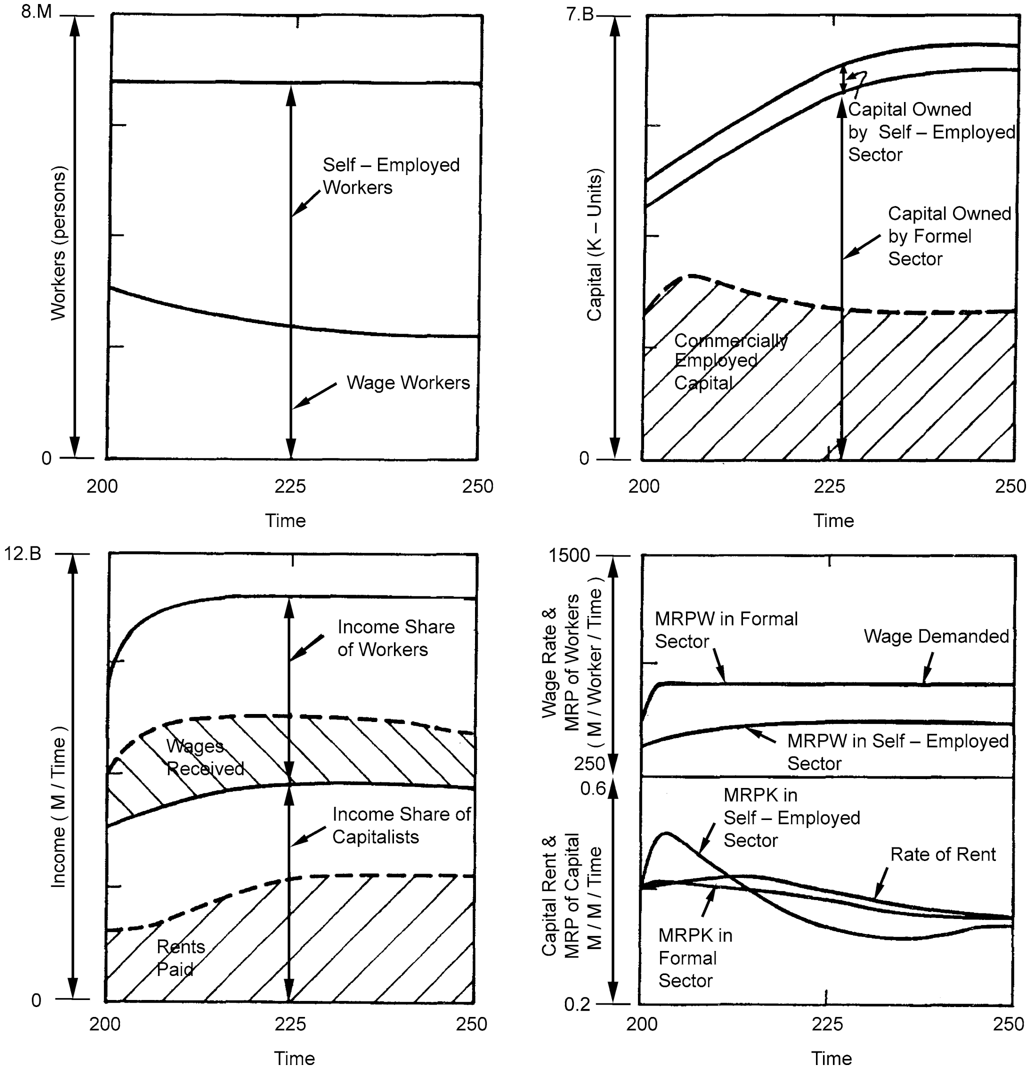
A functional economic system must incorporate the mechanisms to mobilize the forces of self-interest and entrepreneurship inherent in private ownership of the resources. Yet, it must avoid the conflicts inherent in the inequalities of income and resource ownership that led to the creation of the alternative socialist paradigm, which is devoid of such forces. According to the preceding analysis, the fundamental mechanism which creates the possibility of concentration of resource ownership is the equal protection accorded to the artisan and absentee forms of ownership by the prevailing legal norms. The financial fragmentation of households and the differences in their saving patterns further facilitate the expansion of absentee ownership. Technological differences between the capitalist and self-employed sectors not only make possible the side-by-side existence of the two modes of production, they also exacerbate the dichotomy between ownership of resources and workership. Apparently, the policy agenda for changing resource ownership and income distribution patterns should strive to limit renting and should additionally prevent the development of financial fragmentation and technological differentiation between the commercial and self-employed production modes if the objective is to minimize the conflicts related to income distribution.

Assisting the Poor

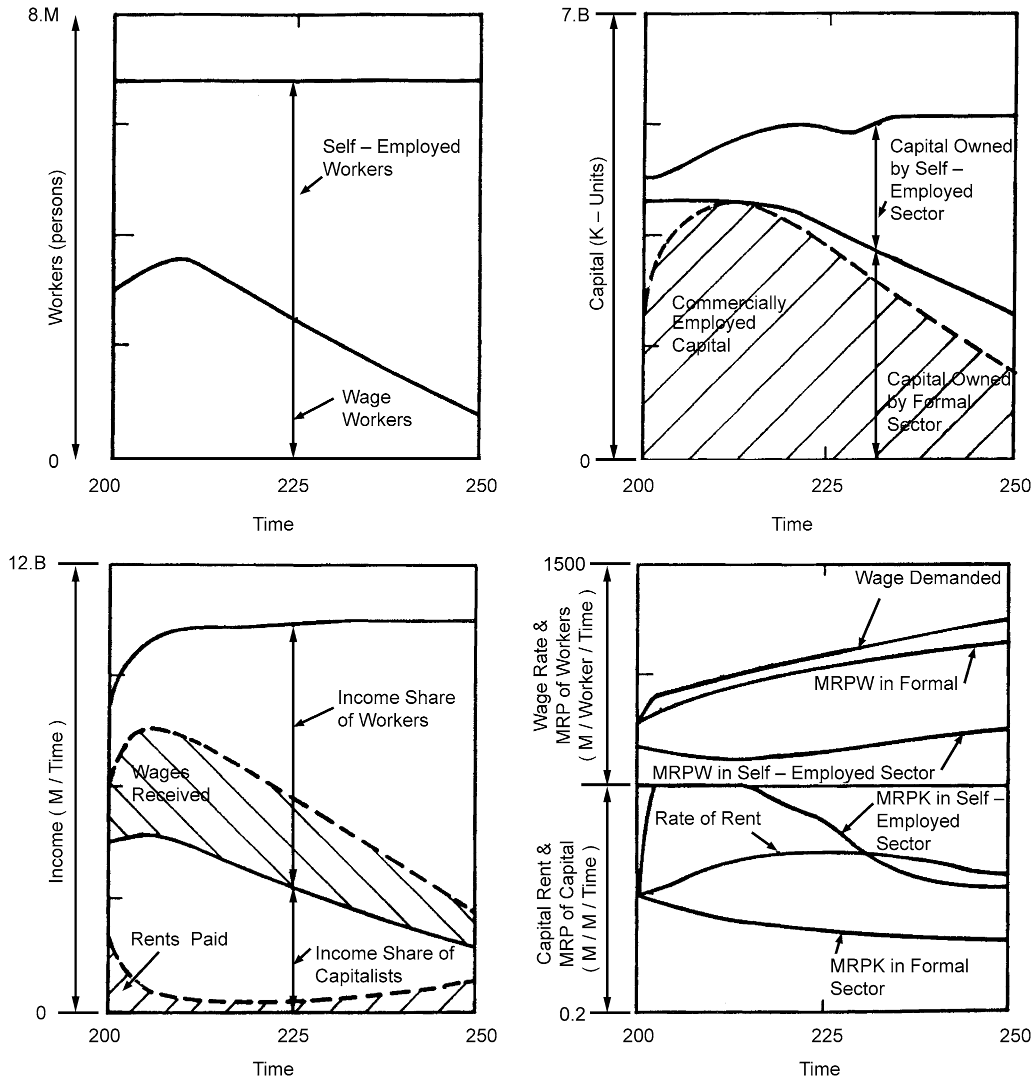
Programs to provide technological, organizational, and financial assistance to the poor have been implemented extensively in the developing countries over the past few decades although they have changed neither income distribution nor wage rate as reflected in many learned writings over these decades as well as the data published by UN and World Bank. This occurred because the increased productivity of the self-employed mode first pushed up wage rate, making renting-out resources more attractive for the capitalist sector than commercial production. However, the consequent decrease in wage payments and increase in rent payments pushed down the income share of the workers, which again suppressed the wage rate. Any efforts to facilitate the small-scale sector to increase its productivity through technological development also failed to affect income distribution since the mechanism of renting allowed the gains of the improved productivity to accrue to the absentee owners of the resources [56]. This experience is verified by the simulation of Fig. 9, which incorporates the policies to improve productivity, creating financial institutions and assisting the self-employed to adopt modern technologies. These policies only increase the size of the self-employed sector without increasing worker income, due to the possibility of separation of the mode of production from the ownership of resources. This indicates that influencing the decision to retain resources in absentee mode for renting out should be the key element of a policy framework to improve income distribution that should alleviate poverty.

Influencing Income Distribution

The cost of owning capital resources in absentee form can be increased by imposing a tax on rent income. The results of implementing this policy, together with the policies of Fig. 9 are shown in Fig. 10. In the face of a tax on rent income, resources which cannot be employed efficiently under the commercial system are offered for sale to the self-employed instead of being leased out to them. Purchase of these resources by the self-employed raises the entitlement of the workers to the income of the economy, which increases the opportunity cost of supplying wage-labor to the commercial sector. This raises wage rate, which makes the commercial mode of production even more uneconomical, unless it is able to apply a superior technology. Such changes spiral in the long run into a transfer of a substantial amount of resources to the self-employed sector. Concomitant efforts to decrease the financial fragmentation of



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 9
Perpetuation of low wage and unequal income distribution resulting from widely used economic development policies



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 10
Changes in wage and income distribution resulting from adding taxation of rent income to the policy package

households and the technological differentiation between the two modes of production, along with improving productivity, further accelerate these changes.

Facilitation of Innovation and Productivity Improvement

Macroeconomic analyses concerning the industrialized countries show that technological innovation is one of the most important sources of growth [9,61]. Studies conducted at the organizational level in the industrialized countries also show that innovations creating technological progress originate largely from small entrepreneurs or from large companies structured in a way to encourage small independent working groups [38,42]. Thus, entrepreneurial activity is often credited with raising productivity through creation of technical and business-related innovations [6]. The high and rising cost of labor in the developed countries, possibly also forces the wage-employers into finding innovative ways of maintaining high labor productivity and continuously striving to improve it.

On the other hand, economic growth has been dominated in the developing countries by relatively large and often highly centralized and vertically integrated companies. Technologies of production have mostly been replanted from the industrialized countries and indigenous innovation and technological development have had a poor track record [1]. These technologies often do not perform as well as at their respective sources, but due to the availability of cheap labor, their inefficient performance still yields comfortable profits; hence little effort is made to improve productivity. There also exist serious limitations on the number of small players as large cross-section of the households in the developing countries lack the resources to effectively participate in any form of entrepreneurial activity [53,58]. Innovation being a probabilistic process, limited participation drastically limits its scope.

There exists a promising institution in most developing countries, however, which has great potential as a focal point of entrepreneurial activity, which has remained dormant for lack of empowerment. This institution is the small family enterprise in the self-employed sector, which may take the form of a shop-house or an artisan manufacturing firm in the urban sector or a peasant farm in the rural sector. It allows participation from all members of the family while also providing the informal small-group organization considered conducive to innovation in many studies. Its members are highly motivated to work hard and assume the risk of enterprise because of their commitment to support the extended family. This enterprise is somewhat similar to the small manufacturing units that created the industrial revolution in England in the early nineteenth century. It has also been observed that the small family enterprise tends to maximize consumption; hence its income significantly affects demand, which creates new marketing opportunities [1,4]. Unfortunately, this enterprise has been systematically suppressed and discriminated against in favor of the large-scale capitalist sector. Even its output remains largely unaccounted for in the national accounting systems of most countries [11,20].

The small family enterprise, variously described as the informal, labor-intensive, traditional, peasant, peripheral and sometimes inefficient sector in the developing countries has been stifled in the first instance by a set of social and legal norms through which the wealth has become concentrated in an absentee ownership mode. Working households are mostly poor and own few assets [19]. The prosperity of these households will not only provide the much-needed financial resources for entrepreneurial activity, their capacity to spend will also create many marketing opportunities for the potential entrepreneur. Thus, influencing income distribution, through the policy framework proposed in the last section, ranks first on the agenda also for developing entrepreneurship and encouraging innovation. Once a significant cross-section of the populace becomes a potential participant in the economic activity, the development of infrastructure and facilitation to manage risk will also appear to be effective instruments to support entrepreneurship and innovation [51]. The rise in the wage rates due to the possibility of alternative self-employment opportunities would, at the same time, force the large commercial enterprise to invest in technological innovation for productivity improvement, which should further improve the efficacy of the overall system.

Conclusion

Both neoclassical and Marxist models of economic growth seem to make restricting assumptions about ownership and mechanisms of wage determination, which are linked with specific time- and geography- related historical evidence. These restricting assumptions give internal consistency and a semblance of sustainability to each model, although they remove both from reality. A failure in the free-market system based on the no-classical model occurs when the invisible hand concentrates ownership of resources in a small minority, suppressing wage rate and creating social conflict due to income inequalities. On the other hand, a failure in the socialist system based on the Marxist model occurs, when the visible hand empowered to act in the public interest stifles entrepreneurial energy while also ignoring public interest in favor of its power interests [48,50].

A behavioral model underlying wage and income distribution has been proposed in this paper, in which the opportunity cost of supplying a unit of labor to the capitalist sector is used as a basis for negotiating a wage. Neither this opportunity cost nor the ownership pattern are taken as given, while the dynamic interaction between the two creates a tendency in the system to generate numerous wage and income distribution patterns, subsuming those postulated in the neo-classical and Marxist theories of economics. The realization of a specific wage and income distribution pattern depends on legal and social norms concerning ownership, renting, the financing of investment and the state of technology.

Private ownership seems to have three forms, commercial, artisan and absentee. Predominance of artisan ownership creates an egalitarian wage and income distribution pattern while a healthy competition between the commercial and artisan firms may release considerable entrepreneurial energy. These functional forms can grow only if the renting of resources can be discouraged. On the other hand, absentee ownership creates a low wage rate and an unequal income distribution, while the growth of this form of ownership is facilitated through the renting mechanism. Potentially, all three ownership forms can exist in an economic system. The problem, therefore, is not to favor or condemn private ownership *per se* as the alternative theories of economics have often advocated, but to understand the reasons behind the development of a particular ownership pattern and identify human motivational factors that would change an existing pattern into a desired one.

The most important reform needed at government level to alleviate poverty is the discouragement of the absentee ownership of capital assets, which would create a wider distribution of wealth. Widespread artisan ownership resulting from this would increase participation in entrepreneurial activity, which would allow adequate performance from the human actors in the system. Such reforms may however not be possible in the authoritarian systems of government pervasive in the developing countries since they must often limit civil rights and public freedoms to sustain power. Hence, the creation of a democratic political system may be a pre-condition to any interventions aimed at poverty alleviation. This, I have discussed elsewhere [50,53,54].

Future Directions

While the market system has often been blamed by the proponents of central planning for leading to concentration of wealth and income among few, it in fact offers a powerful means for redistributing income if the process of concentration is carefully understood and an intervention designed on the basis of this understanding. In fact, all economic systems can be reformed to alleviate the dysfunctional tendencies they are believed to have, provided the circular relationships creating such dysfunctions can be understood which should be the first objective of policy design for economic development.

Contrary to this position, economic development has often viewed developmental problems as pre-existing conditions, which must be changed through external intervention. Poverty, Food shortage, poor social services and human resources development infrastructure, technological backwardness, low productivity, resource depletion, environmental degradation and poor governance are cases in point. In all such cases, the starting point for a policy search is the acceptance of a snapshot of the existing conditions. A developmental policy is then constructed as a well-intended measure that should improve existing conditions. Experience shows, however, that policies implemented with such a perspective not only give unreliable performance, they also create unintended consequences. This happens because the causes leading to the existing conditions and their future projections are not adequately understood. The well-intentioned policies addressing problem symptoms only create ad hoc changes, which are often overcome by the system's reactions.

Table 1 collects three key developmental problems, poverty, food shortage and social unrest, and the broad policies implemented over the past several decades to address them. These problems have, however, continued to persist or even become worse.

The policy response for overcoming poverty was to foster economic growth so aggregate income could be increased; that for creating food security was intensive agri-

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Table 1
Developmental problems, policies implemented to address them and unintended consequences experienced

Initially perceived problems	Policies implemented	Unintended consequences
Poverty	Economic growth capital formation sectoral development technology transfer external trade	Low productivity indebtedness natural resources depletion environmental degradation continuing/increased poverty
Food shortage	Intensive agriculture land development irrigation fertilizer application use of new seeds	Land degradation depletion of water aquifers vulnerability to crop failure population growth continuing/increased vulnerability to food shortage
Social unrest	Spending on internal security and defense infrastructure limiting civil rights	Poor social services poor economic infrastructure authoritarian governance insurgence continuing/increased social unrest

culture so more food could be produced; and for containing social unrest, the broad prescription is to strengthen internal security and defense infrastructure so public could be protected from social unrest. The unintended consequences of these policies are many, but in most instances, they include a continuation or worsening of the existing problems.

Thus, poverty and income differentials between rich and poor have in fact shown a steady rise, which is also accompanied by unprecedented debt burdens and extensive depletion of natural resources and degradation of environment. Food shortages have continued but are now accompanied also by land degradation, depletion of water aquifers, the threat of large-scale crop failure due to a reduction in crop diversity and a tremendous growth in population. Social unrest has often intensified together with appearance of organized insurgence burgeoning expenditures on internal security and defense, which has stifled development of social services and human resources and have created authoritarian governments with little commitment to public welfare.

The unintended consequences are often more complex than the initial problems and have lately drawn concerns at the global level, but whether an outside hand at the global level would alleviate them is questionable. This is evident from the failure to formulate and enforce global public policy in spite of active participation by national governments, global agencies like the UN, the World Bank, the World Trade Organization, and advocacy networks sometimes referred to as the civil society. This failure can largely be attributed to the lack of a clear understanding of the roles of the actors who precipitated those problems and whose motivations must be influenced to turn the tide.

Thus, development planning must adopt a problem solving approach in a mathematical sense if it is to achieve sustainable solutions. In this approach, a problem must be defined as an internal behavioral tendency and not as a snap shot of existing conditions. It may represent a set of patterns, a series of trends or a set of existing conditions that appear either to characterize a system or to be resilient to policy intervention. In other words, an end condition by itself must not be seen as a problem definition. The complex pattern of change implicit in the time paths preceding this end condition would, on the other hand, represent a problem. The solution to a recognized problem should be a solution in a mathematical sense, which is analogous to creating an understanding of the underlying causes of a delineated pattern. A development policy should then be perceived as a change in the decision rules that would change a problematic pattern to an acceptable one. Such a problem solving approach can be implemented with advantage using system dynamics modeling process that entails building and experimenting with computer models of problems, provided of course a succinct problem definition has first been created.

Appendix

Model Description

Wage rate *WR* is assumed to adjust over period *WRAT* towards indicated wage rate *IWR*.

$$\mathrm{d}/\mathrm{d}t[WR] = (IWR - WR)/WRAT \tag{1}$$

IWR depends on the wage-bargaining position of the workers, which is determined by their opportunity cost of accepting wage-employment. It is assumed that the opportunity cost of transferring a self-employed worker to wage-work is zero when wage offered is equal to the current consumption expenditure per worker averaged over the whole workforce.

$$IWR = [(R_\mathrm{s} * (1 - SP_\mathrm{s}) + (AS_\mathrm{s}/LAS))/TW] \,, \tag{2}$$

where R_s, SP_s and $\mathrm{AS_s}$ are, respectively, income share, saving propensity and accumulated unspent savings of the self-employed sector. *LAS* and *TW* are, respectively, life of accumulated unspent savings and total workforce. Subscripts s and f designate, respectively, self-employed and capitalist sectors.

Ownership of land and capital as well as contribution to labor are the bases for claim to income while absentee ownership is possible through leasing arrangements. Thus, R_s is computed by adding together the value of output produced by the self-employed sector VQ_s and the wage payments received by the wage-workers W_f, and subtracting from the sum the rent payments made to the absentee owners. R_f is given by adding together the value of output produced by the capitalist sector VQ_f and the rent payments it receives from the self-employed sector, and subtracting from the sum the wage-payments it makes.

$$R_\mathrm{s} = VQ_\mathrm{s} + WR * W_\mathrm{f} - LR * RL - KR * RK \,, \tag{3}$$

$$R_\mathrm{f} = VQ_\mathrm{f} - WR * W_\mathrm{f} + LR * RL + KR * RK \,, \tag{4}$$

where *LR*, *RL*, *KR*, and *RK*, are, respectively, land rent, rented land, capital rent, and rented capital.

KR and *LR* depend, respectively, on the long-term averages of the marginal revenue products of capital and land (*AMRPK* and *AMRPL*) in the economy, and the demand for renting capital and land (*RKD* and *RLD*) as compared with the supply of rentable assets (*RK* and *RL*). The demand for renting, in turn, depends on the lack of ownership of adequate resources for productively employing the workers in the self-employed sector.

$$KR = AMRPK * f_1[RKD/RK]\,; \quad f_1' > 0 \tag{5}$$

$$RKD = DKE_s - KO_s\,. \tag{6}$$

Where DKE_s is desired capital to be employed in the self-employed sector and KO_s is capital owned by it. Land rent *LR* and demand for renting land *RLD* are determined similarly.

The saving propensity of all households in not uniform. Since capitalist households associated with the capitalist sector receive incomes which are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the worker households depends on their need to save for supporting investment for self-employment and on how their absolute level of income compares with their inflexible consumption. Thus, SP_s in the model is determined by the utility of investment in the self-employed sector arising from a comparison of worker productivity in the sector with the wage rate in the capitalist sector, and the rent burden of this sector compared with the factor contribution to its income from land and capital.

$$\begin{aligned} SP_s = \mu * f_2[MRPW_s/WR] * f_3[(&LR * RL + KR * RK) \\ &/(VQ_s - MRPW_s * W_s)]\,, \end{aligned} \tag{7}$$

$$SP_f = \mu\,, \tag{8}$$

where $f_2' > 0$, $f_3' < 0$, μ is a constant, and MRPW is marginal revenue product of workers.

AS represent the balance of unspent savings, which determine the availability of liquid cash resources for purchase of assets. AS are consumed over their life LAS whether or not any investment expenditure occurs.

$$\begin{aligned} &\mathrm{d}/\mathrm{d}t[AS_i] \\ &\quad = R_i * SP_i - AS_i/LAS - LA_i * PL - \sum_j KA_i^j * GPL\,; \\ &\qquad\qquad\qquad\qquad i = \mathrm{s,f}\,; \quad j = \mathrm{m,t}\,, \end{aligned} \tag{9}$$

where *LA*, *PL*, *KA*, and *GPL* are, respectively, land acquisitions, price of land, capital acquisitions, and general price level. Subscript *i* refers to any of the two sectors, self-employed (s) and capitalist (f), and superscript *j* to the type of capital, modern (m) or traditional (t).

W_f is assumed to adjust towards indicated workers IW_f given by desired workers DW_f and total workforce *TW*. *TW* is assumed to be fixed, although, relaxing this assumption does not alter the conclusions of this paper. All workers who are not wage-employed must be accommodated in self-employment. Thus W_s represents the remaining workers in the economy.

$$\mathrm{d}/\mathrm{d}t[W_f] = (IW_f - W_f)/WAT \tag{10}$$

$$IW_f = TW * f_4(DW_f/TW) \tag{11}$$

$$W_s = TW - W_f \tag{12}$$

where $1 \geq f_4 \geq 0$, and $f_4' > 0$. WAT is worker adjustment time.

The desired workers in each sector DW_i is determined by equating wage rate with the marginal revenue product of workers. A modified Cobb–Douglas type production function is used.

$$DW_i = E_i^w * VQ_i/WR\,, \tag{13}$$

where E_i^w is the elasticity of production of workers in a sector.

Land and capital owned by the capitalist sector (LO_f and KO_f) are allocated to commercial production (KE_f and LE_f) and renting (*RK* and *RL*) activities depending on the desired levels of these factors in each activity. Thus,

$$RK = (DRK/(DRK + DKE_f)) * KO_f \tag{14}$$

$$RL = (DRL/(DRL + DLE_f)) * LO_f \tag{15}$$

$$KE_f = KO_f - RK \tag{16}$$

$$LE_f = LO_f - RL \tag{17}$$

Capital and land employed by the self-employed sector consist of these production factors owned by them and those rented from the capitalist sector.

$$KE_s = Ko_s + RK \tag{18}$$

$$LE_s = Lo_s + RL\,. \tag{19}$$

Desired capital and land to be employed in any sector (DKE_i and DLE_i) are determined on the basis of economic criteria.

$$\mathrm{d}/\mathrm{d}t(DKE_i)/KE_i = f_6[MRPK_i/MFCK] \tag{20}$$

$$\mathrm{d}/\mathrm{d}t(DLE_i)/LE_i = f_5[MRPL_i/MFCL]\,, \tag{21}$$

where f_5' and $f_6' > 0$. $MRPL_i$ and $MRPK_i$ are respectively marginal revenue products of land and capital in a sector, and *MFCL* AND *MFCK* are respectively marginal factor costs of land and capital.

$$MRPL_i = (E_i^{\mathrm{l}} * VQ_i/LE_i) \tag{22}$$

$$MRPK_i = (E_i^{\mathrm{k}} * VQ_i/KE_i) \tag{23}$$

$$MFCL = PL * IR \tag{24}$$

$$MFCK = IR + (1/LK) * GPL\,, \tag{25}$$

where E_i^{l} and E_i^{k} are, respectively, elasticities of production of land and capital in a sector. *PL* is price of land, *IR* is exogenously defined interest rate, *LK* is life of capital and *GPL* is general price level.

Changes in the quantities of capital and land desired to be rented out (*DRK* and *DRL*) depend on their respective rents *KR* and *LR* compared with their marginal factor costs *MFCK* and *MFCL*.

$$\mathrm{d}/\mathrm{d}t[DRK]/RK = f_7[KR/MFCK]\,; \quad f_7' > 0 \tag{26}$$

$$\mathrm{d}/\mathrm{d}t[DRL]/RL = f_8[LR/MFCL]\,; \quad f_8' > 0\,. \tag{27}$$

The value of output produced by each sector is given by the product of the quantity it produces Q_i and the general price level GPL.

$$VQ_i = Q_i * GPL \tag{28}$$

$$Q_i = A_i * K_i^{E^{\mathrm{ki}}} * L_i^{E^{\mathrm{li}}} * W_i^{E^{\mathrm{wi}}}\,, \tag{29}$$

where K_i, L_i, and W_i represent capital, land and workers employed by a sector. A_i represent technology constants, which increase with the use of modern capital.

$$A_i = \mathring{A} * f_9\big[K_i^m/(K_i^t + K_i^m)\big]\,, \tag{30}$$

where $f_9' > 0$ and $\mathring{A}$ is a scaling factor based on initial conditions of inputs and output of the production process.

Ownership is legally protected and the financial market is fragmented by households. Thus, purchase of any productive assets must be self-financed by each sector through cash payments. Land ownership LO_i of each sector changes through acquisitions LA_i from each other. Each sector bids for the available land on the basis of economic criteria, its current holdings, and the sector's liquidity.

$$LA_i = \mathrm{d}/\mathrm{d}t[LO_i] \tag{31}$$

$$LO_i = \left(DLO_i/\sum_i DLO_i\right) * TL\,, \tag{32}$$

where DLO_i is desired land ownership in a sector and *TL* is total land which is fixed,

$$DLO_i = LO_i * f_6[MRPL_i/MFCL] * f_{11}[CA_i]\,, \tag{33}$$

where $f_{11}'[CA_i]$ is > 0, and CA_i is cash adequacy of a sector.

Cash adequacy of a sector CA_i is given by the ratio of its accumulated unspent savings to the desired savings. The latter is computed by multiplying cash needed to finance investment and the traditional rate of consumption of savings in the sector by cash coverage *CC*.

$$CA_i = AS_i/\left(\left((AS_i/LAS) + (LA_i * PL) + \Big(\sum_j KA_{ij} * GPL\Big)\right) * CC\right). \tag{34}$$

Capital ownership in a sector $KO_i = KO_i^t + KO_i^m$ changes through acquisitions KA_i^j and decay. Although there is a preference for modern capital, its acquisition KA_i^m depends on the ability to accommodate the technology represented by it. Inventory availability of each type of capital KIA^j also limits its purchases.

$$\mathrm{d}/\mathrm{d}t[KO_i] = \sum_j KA_i^j - KO_i/LK\,, \tag{35}$$

$$KA_i^j = DKA_i^j * KIA^j\,, \tag{36}$$

$$DKA_i^m = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * TCF_i\,, \tag{37}$$

$$DKA_i^t = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * (1 - TCF_i)\,, \tag{38}$$

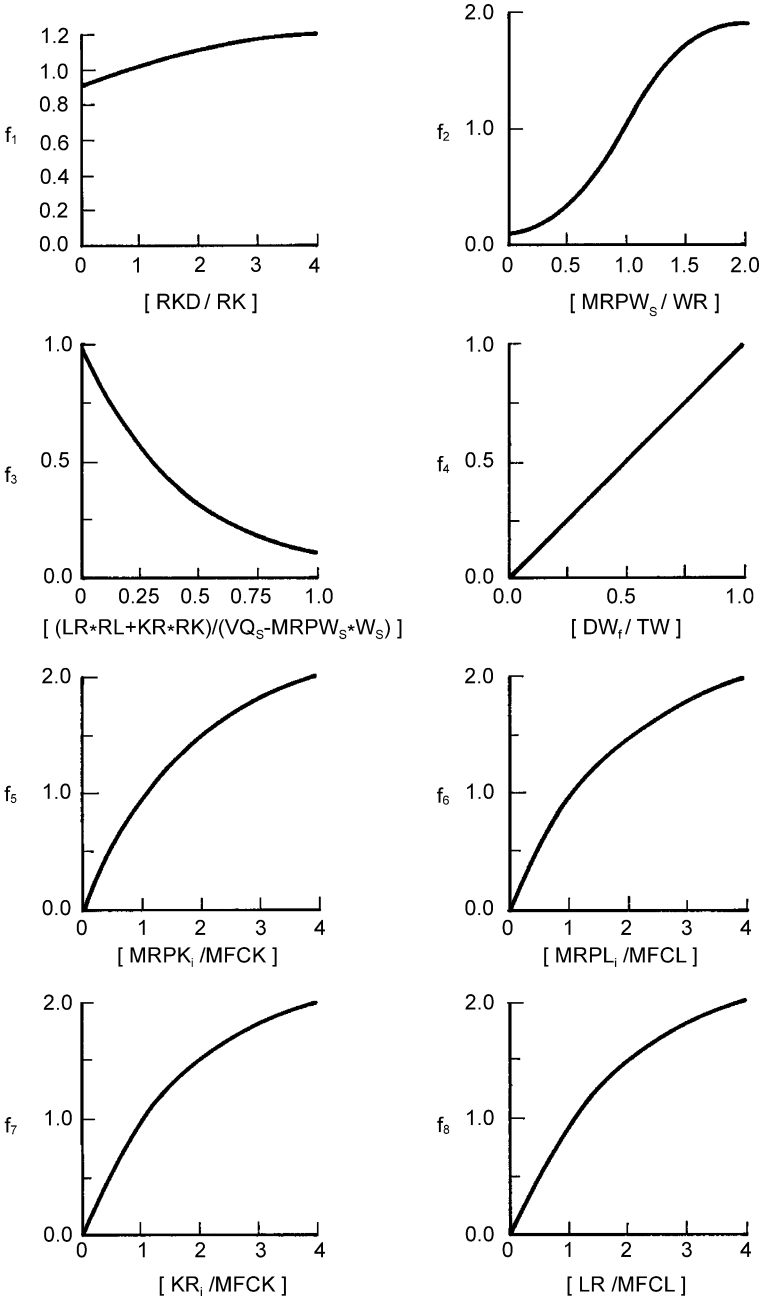
where DKA_i are desired capital acquisitions, $f_{11}' \geq 0$, and *LK* is life of capital. TCF_i represent exogenously defined technological capability. $0 < TCF_i < 1$.

$$KIA^j = f_{12}\left[KI^j/\left(\sum_i DKA_i^j\right) * KIC\right]\,, \tag{39}$$

where $0 \leqslant f_{12} \leqslant 1$, $f_{12}' > 0$, and *KIC* is capital inventory coverage

$$\mathrm{d}/\mathrm{d}t[KI^j] = KQ^j - \sum_i KA_i^j\,, \tag{40}$$

where KQ^j represent supply of capital. KQ^m is imported, while KQ^t is created within the economy by allocating a part of the capacity to its production.



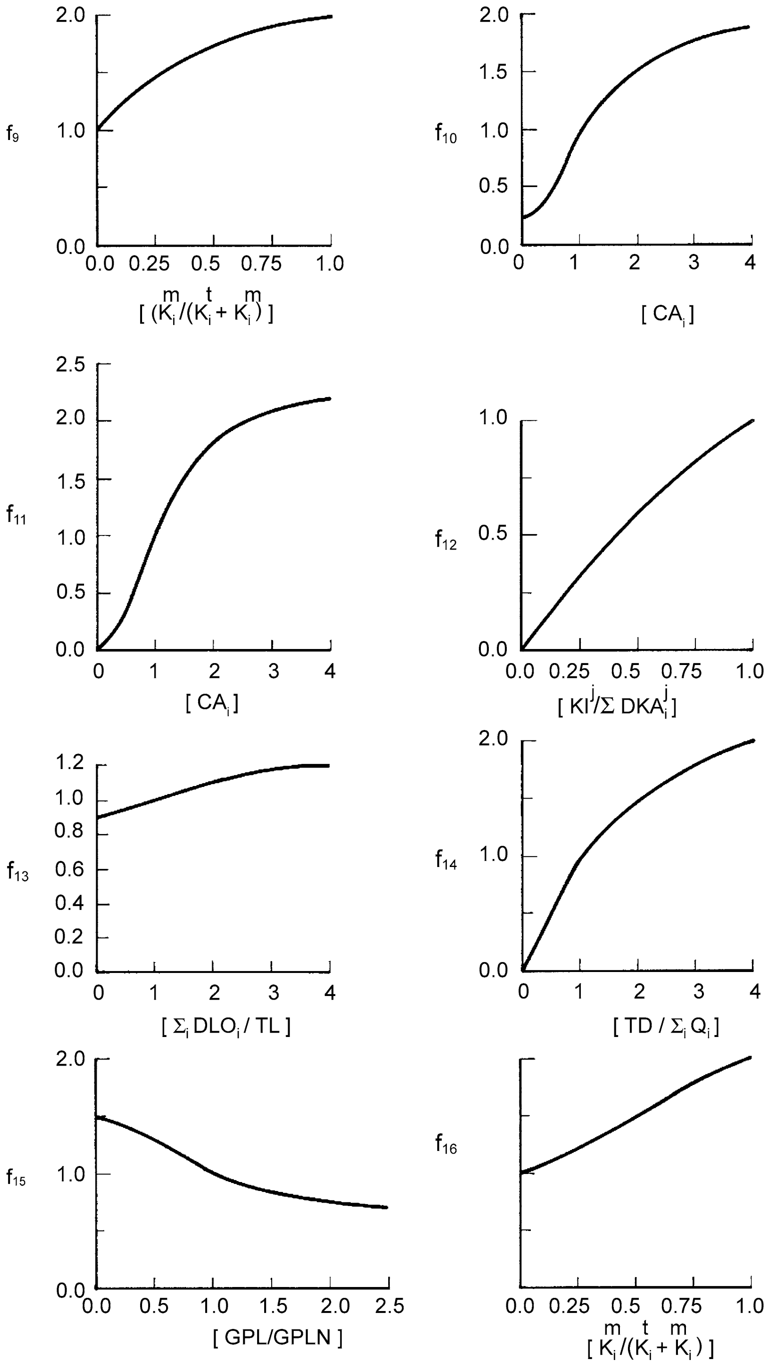
Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 11
Behavioral relationships f_1 through f_8

$$KQ^t = \sum Q_i * \left(\sum_i DKA_i^t / TD \right) . \qquad (41)$$

The price of land *PL* is assumed to adjust towards indicated price of land *IPL* which is given by the economy-wide average of the marginal revenue product of land *AMRPL*, interest rate *IR* and the desired land ownership in each sector DLO_i

$$\mathrm{d}/\mathrm{d}t[PL] = (IPL - PL)/LPAT \, , \qquad (42)$$

$$IPL = (AMRPL/IR) * f_{13} \left[\sum_i DLO_i / TL \right] ; \quad \text{where} \quad f'_{13} > 0 \, . \qquad (43)$$



Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 12
Behavioral relationships f_9 through f_{16}

General price level GPL is determined by supply and demand considerations.

$$d/dt[GPL] = GPLN * f_{14}[TD/\sum_i Q_i] \qquad (44)$$

where $f'_{14} > 0$. *GPLN* is normal value of *GPL* and *TD* is total demand for goods and services to be produced within the economy. *TD* is given by adding up non-food consumption C_i, traditional capital acquisition KA_i^t and production of traditional capital for inventory, food demand *FD* and government spending G which is equal to taxes, if any, collected.

$$\begin{aligned} TD = {} & \sum C_i + \sum_i DKA_i^t \\ & + \left(\left(KIC * \sum_i DKA_i^t - KI^j \right) / IAT \right) \\ & + FD + G \,, \end{aligned} \tag{45}$$

$$\mathrm{d}/\mathrm{d}t(C_i) = \frac{[(((R_i * (1 - SP_i) + AS_i/LAS)/GPL) * FNFC_i) - C_i]}{CAT} \,, \tag{46}$$

where *IAT* is inventory adjustment time, $FNFC_i$ fraction non-food consumption, and *CAT* is consumption adjustment time. Food demand *FD* is given by multiplying population *P* with normal per capita food demand *NFPCD* and a function f_{15} representing a weak influence of price.

$$FD = P * NFPCD * f_{15}[GPL/GPLN] \,, \tag{47}$$

where $f'_{15} < 0$ and *P* bears a fixed proportion with total workforce *TW*.

The elasticity of production of land E_i^{l} is assumed to be constant as is suggested by empirical evidence concerning agricultural economies [Strout 1978, Heady and Dillon 1961]. Elasticity of production of capital E_i^{k} depends on the technology of production, which is determined by the proportions of traditional and modern capital employed. Since constant returns to scale are assumed, E_i^{w} is given by (47).

$$E_i^{\mathrm{k}} = f_{16}[K_i^m/(K_i^t + K_i^m)] \,; \quad f'_{16} > 0 \,, \tag{48}$$

$$E_i^{\mathrm{w}} = 1 - E_i^{\mathrm{k}} - E_i^{\mathrm{l}} \,. \tag{49}$$

Behavioral Relationships

Sixteen behavioral relationships $[f_1 \cdots f_{16}]$ have been incorporated into the model. The slope characteristics of these relationships have already been described in above equations. The graphical forms of the functions representing these relationships are shown in Figs. 11 and 12 placed below. General considerations for specifying such relationships are discussed in [63].

Bibliography

Primary Literature

1. APO (1985) Improving productivity through macro-micro linkage. Survey and Symposium Report. Tokyo: Asian Productivity Organization
2. Applebaum E (1979) The labor market. In: Eichner A (ed) A Guide to Post-Keynesian Economics. ME Sharpe, White Plains, New York
3. Averitt RT (1968) The dual economy: the dynamics of american industry structure. Norton, New York
4. Bardhan PK (1973) A model of growth in a dual agrarian economy. In: Bhagwati G, Eckus R (eds) Development and planning: essays in honor of Paul Rosenstein-Roden. George Allen and Unwin Ltd, New York
5. Barro RJ (1997) Macroeconomics, 5th edn. MIT Press, Cambridge
6. Baumol WJ (1988) Is entrepreneurship always productive? J Dev Plan 18:85–94
7. Boeke JE (1947) Dualist economics. Oriental economics. Institute of Pacific Relations, New York
8. Cornwall J (1978) Growth and stability in a mature economy. John Wiley, London
9. Denison EF (1974) Accounting for United States economic growth 1929–1969. Brookings Institution, Washington
10. Eichner A, Kregel J (1975) An essay on post-Keynesian theory: a new paradigm in economics. J Econ Lit 13(4):1293–1314
11. Eisner R (1989) Divergences of measurement theory and some implications for economic policy. Am Econ Rev 79(1):1–13
12. Fie JC, Ranis G (1966) Agrarianism, dualism, and economic development. In: Adelman I, Thorbecke E (eds) The Theory and Design of Economic Development. Johns Hopkins Press, Baltimore
13. Forrester JW (1979) Macrobehavior from microstructure, In: Karmany NM, Day R (eds) Economic issues of the eighties. Johns Hopkins University Press, Baltimore
14. Forrester JW (1987) Lessons from system dynamics modelling. Syst Dyn Rev 3(2):136–149
15. Galbraith JK (1979) The nature of mass poverty. Harvard University Press, Cambridge
16. Gordon DM (1972) Economic theories of poverty and underemployment. DC Heath, Lexington
17. Graham AK (1977) Principles of the relationships between structure and behavior of dynamic systems. Ph D Thesis. MIT, Cambridge
18. Griffin K, Ghose AK (1979) Growth and impoverishment in rural areas of Asia. World Dev 7(4/5):361–384
19. Griffin K, Khan AR (1978) Poverty in the third world: ugly facts and fancy models. World Dev 6(3):295–304
20. Hicks J (1940) The valuation of the social income. Economica 7(May):163–172
21. Higgins B (1959) Economic development. Norton, New York
22. Hirshliefer J (1976) Price theory and applications. Prentice Hall, Englewood Cliffs
23. Kaldor N (1966) Marginal productivity and the macro-economic theories of distribution. Rev Econ Stud 33:309–319
24. Kaldor N (1969) Alternative theories of distribution. In: Stiglitz J, Ozawa H (eds) Readings in modern theories of economic growth. MIT Press, Cambridge

25. Kalecki M (1965) Theory of economic dynamics, revised edn. Allen and Unwin, London
26. Kalecki M (1971) Selected essays on dynamics of capitalist economy. Cambridge Univ Press, London
27. Kindelberger C, Herrick B (1977) Economic development, 3rd edn. McGraw Hill, New York
28. Leontief W (1977) Theoretical assumptions and non-observable facts. In: Leontief W (ed) Essays in Economics, vol II. ME Sharpe, White Plains
29. Lewis WA (1958) Economic development with unlimited supply of labor. In: Agarwala I, Singh SP (eds) The Economics of Underdevelopment. Oxford University Press, London
30. Lipton M (1977) Why poor people stay poor. Harvard University Press, Cambridge
31. Marglin SA (1984) Growth, distribution and prices. Harvard Univ Press, Cambridge
32. Marx K (1891) Capital. International Publishers, New York (Reprinted)
33. McKinnon RI (1973) Money and capital in economic development. The Brookings Institution, New York
34. Minsky H (1975) John Maynard Keynes. Columbia University Press, New York
35. Mukhia H (1981) Was there feudalism in indian history. J Peasant Stud 8(3):273–310
36. Myrdal G (1957) Economic theory and under-developed regions. Gerald Duckworth Ltd, London
37. Pack SJ (1985) Reconstructing Marxian economics. Praeger, New York
38. Quinn JB (1985) Managing innovation: controlled chaos. Harvard Bus Rev 85(3):73–84
39. Ricardo D (1817) Principles of political economy and taxation, Reprint 1926. Everyman, London
40. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
41. Riech M, Gordon D, Edwards R (1973) A theory of labor market segmentation. Am Econ Rev 63:359–365
42. Roberts EB (1991) Entrepreneurs in high technology, lessons from MIT and beyond. Oxford University Press, New York
43. Robinson J (1955) Marx, Marshal and Keynes: three views of capitalism. Occasional Paper No. 9. Delhi School of Economics, Delhi
44. Robinson J (1969) The theory of value reconsidered. Aust Econ Pap June 8:13–19
45. Robinson J (1978) Contributions to modern economics. Basil Blackwell, Oxford
46. Robinson J (1979) Aspects of development and underdevelopment. Cambridge University Press, London
47. Roulet HM (1976) The Historical context of Pakistan's rural agriculture. In: Stevens RD et al (eds) Rural development in Bangladesh and Pakistan. Hawaii University Press, Honolulu
48. Rydenfelt S (1983) A pattern for failure. Socialist economies in crisis. Harcourt Brace Jovanavich, New York
49. Saeed K (1988) Wage determination, income distribution and the design of change. Behav Sci 33(3):161–186
50. Saeed K (1990) Government support for economic agendas in developing countries. World Dev 18(6):758–801
51. Saeed K (1991) Entrepreneurship and innovation in developing countries: basic stimulants, organizational factors and hygienes. Proceedings of Academy of International Business Conference. Singapore, National University of Singapore
52. Saeed K (1992) Slicing a complex problem for system dynamics modelling. Syst Dyn Rev 8(3):251–261
53. Saeed K (1994) Development planning and policy design: a system dynamics approach. Foreword by Meadows DL. Ashgate/Avebury Books, Aldershot
54. Saeed K (2002) A pervasive duality in economic systems: implications for development planning. In: "Systems dynamics: systemic feedback modeling for policy analysis". In: Encyclopedia of life support systems (EOLSS). EOLSS Publishers, Oxford
55. Saeed K (2005) Limits to growth in classical economics. 23rd International Conference of System Dynamics Society, Boston
56. Saeed K, Prankprakma P (1997) Technological development in a dual economy: alternative policy levers for economic development. World Dev 25(5):695–712
57. Sen AK (1966) Peasants and dualism with or without surplus labor. J Polit Econ 74(5):425–450
58. Sen AK (1999) Development as freedom. Oxford University Press, Oxford
59. Simon HA (1982) Models of bounded rationality. MIT Press, Cambridge
60. Skinner A (ed) (1974) Adam Smith: The wealth of nations. Pelican Books, Baltimore
61. Solow R (1988) Growth theory and after. Am Econ Rev 78(3):307–317
62. Sraffa P (1960) Production of commodities by means of commodities. Cambridge University Press, Cambridge
63. Sterman J (2000) Business dynamics. McGraw Hill, Irwin
64. Streeten P (1975) The limits of development studies. 32nd Montague Burton Lecture on International Relations. Leads University Press, Leeds
65. Takahashi Y et al (1970) Control and dynamic systems. Addison-Wesley, Reading
66. Weintraub S (1956) A macro-economic approach to theory of wages. Am Econ Rev. 46(Dec):835–856

Books and Reviews

Atkinson G (2004) Common ground for institutional economics and system dynamics modeling. Syst Dyn Rev 20(4):275–286

Ford A (1999) Modeling the environment. Island Press, Washington DC

Forrester JW (1989) The system dynamics national model: macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin

Forrester N (1982) A Dynamic synthesis of basic macroeconomic theory: implications for stabilization policy analysis. Ph D Dissertation, MIT

Forrester N (1982) The life cycle of economic development. Pegasus Communications, Walthum

Hines J (1987) Essays in behavioral economic modeling. Ph D Dissertation, MIT, Cambridge

Mass NJ (1976) Economic cycles, an analysis of the underlying causes. Pegasus Communications, Walthum

Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. J Econ Issues 37(1):133–173

Randers J (1980) Elements of system dynamics method. Pegasus Communications, Walthum

Richardson GP (1996) Modeling for management: simulation in support of systems thinking. In: Richardson GP (ed) The international library of management. Dartmouth Publishing Company, Aldershot

Saeed K (1980) Rural development and income distribution, the case of pakistan. Ph D Dissertation. MIT, Cambridge

Saeed K (1994) Development planning and policy design: a system dynamics approach. Ashgate/Avebury Books, Aldershot

Saeed K (1998) Towards sustainable development, 2nd edn: Essays on System Analysis of National Policy. Ashgate Publishing Company, Aldershot

Saeed K (2002) System dynamics: a learning and problem solving approach to development policy. Glob Bus Econ Rev 4(1):81–105

Saeed K (2003) Articulating developmental problems for policy intervention: A system dynamics modeling approach. Simul Gaming 34(3):409–436

Saeed K (2003) Land Use and Food Security – The green revolution and beyond. In: Najam A (ed) Environment, development and human security, perspectives from south asia. University Press of America, Lanham

Saeed K (2004) Designing an environmental mitigation banking institution for linking the size of economic activity to environmental capacity. J Econ Issues 38(4):909–937

Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin McGraw Hill, Boston

Wolstenholme E (1990) System enquiry, a system dynamics approach. John Wiley, Chichester

Econometrics: Models of Regime Changes

JEREMY PIGER[1]
Department of Economics,
University of Oregon, Eugene, USA

Article Outline

Glossary

Filtered probability of a regime The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period *t*, conditional on observing sample information up to period *t*.

Gibbs sampler An algorithm to generate a sequence of samples from the joint probability distribution of a group of random variables by repeatedly sampling from the full set of conditional distributions for the random variables.

Markov chain A process that consists of a finite number of states, or regimes, where the probability of moving to a future state conditional on the present state is independent of past states.

Markov-switching model A regime-switching model in which the shifts between regimes evolve according to an unobserved Markov chain.

Regime-Switching Model A parametric model of a time series in which parameters are allowed to take on different values in each of some fixed number of regimes.

Smooth transition threshold model A threshold model in which the effect of a regime shift on model parameters is phased in gradually, rather than occurring abruptly.

Smoothed probability of a regime The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period *t*, conditional on observing all sample information.

Threshold model A regime-switching model in which the shifts between regimes are triggered by the level of an observed economic variable in relation to an unobserved threshold.

Time-varying transition probability A transition probability for a Markov chain that is allowed to vary depending on the outcome of observed information.

Transition probability The probability that a Markov chain will move from state *j* to state *i*.

Definition of the Subject

Regime-switching models are time-series models in which parameters are allowed to take on different values in each of some fixed number of "regimes." A stochastic process assumed to have generated the regime shifts is included as part of the model, which allows for model-based forecasts that incorporate the possibility of future regime shifts. In certain special situations the regime in operation at any point in time is directly observable. More generally the regime is unobserved, and the researcher must conduct inference about which regime the process was in at past points in time. The primary use of these models in the applied econometrics literature has been to describe changes in the dynamic behavior of macroeconomic and financial time series.

Regime-switching models can be usefully divided into two categories: "threshold" models and "Markov-switching" models. The primary difference between these approaches is in how the evolution of the state process is modeled. Threshold models, introduced by Tong [91], assume that regime shifts are triggered by the level of observed variables in relation to an unobserved threshold. Markov-switching models, introduced to econometrics by [16,39,41], assume that the regime shifts evolve according to a Markov chain.

Regime-switching models have become an enormously popular modeling tool for applied work. Of particular note are regime-switching models of measures of economic output, such as real Gross Domestic Product (GDP), which have been used to model and identify the phases of the business cycle. Examples of such models include [3,7,41,57,60,61,73,75,77,90,93]. A sampling of other applications include modeling regime shifts in inflation and interest rates [2,25,34], high and low volatility regimes in equity returns [23,46,48,92], shifts in the Federal Reserve's policy"rule" [55,83], and time variation in the response of economic output to monetary policy actions [35,53,69,81].

[1] I am grateful to Jim Hamilton and Bruce Mizrach for comments on an earlier draft.

Introduction

There is substantial interest in modeling the dynamic behavior of macroeconomic and financial quantities observed over time. A challenge for this analysis is that these time series likely undergo changes in their behavior over reasonably long sample periods. This change may occur in the form of a "structural break", in which there is a shift in the behavior of the time series due to some permanent change in the economy's structure. Alternatively, the change in behavior might be temporary, as in the case of wars or "pathological" macroeconomic episodes such as economic depressions, hyperinflations, or financial crises. Finally, such shifts might be both temporary and recurrent, in that the behavior of the time series might cycle between regimes. For example, early students of the business cycle argued that the behavior of economic variables changed dramatically in business cycle expansions vs. recessions.

The potential for shifts in the behavior of economic time series means that constant parameter time series models might be inadequate for describing their evolution. As a result, recent decades have seen extensive interest in econometric models designed to incorporate parameter variation. One approach to describing this variation, denoted a "regime-switching" model in the following, is to allow the parameters of the model to take on different values in each of some fixed number of regimes, where, in general, the regime in operation at any point in time is unobserved by the econometrician. However, the process that determines the arrival of new regimes is assumed known, and is incorporated into the stochastic structure of the model. This allows the econometrician to draw inference about the regime that is in operation at any point in time, as well as form forecasts of which regimes are most likely in the future.

Applications of regime-switching models are usually motivated by economic phenomena that appear to involve cycling between recurrent regimes. For example, regime-switching models have been used to investigate the cycling of the economy between business cycle phases (expansion and recession), "bull" and"bear" markets in equity returns, and high and low volatility regimes in asset prices. However, regime switching models need not be restricted to parameter movement across recurrent regimes. In particular, the regimes might be non-recurrent, in which case the models can capture permanent"structural breaks" in model parameters.

There are a number of formulations of regime-switching time-series models in the recent literature, which can be usefully divided into two broad approaches. The first models regime change as arising from the observed behavior of the level of an economic variable in relation to some threshold value. These "threshold" models were first introduced by Tong [91], and are surveyed by [78]. The second models regime change as arising from the outcome of an unobserved, discrete, random variable, which is assumed to follow a Markov process. These models, commonly referred to as "Markov-switching" models, were introduced in econometrics by [16,39], and became popular for applied work following the seminal contribution of Hamilton [41]. Hamilton and Raj [47] and Hamilton [44] provide surveys of Markov-switching models, while Hamilton [43] and Kim and Nelson [62] provide textbook treatments.

There are by now a number of empirical applications of regime-switching models that establish their empirical relevance over constant parameter alternatives. In particular, a large amount of literature has evaluated the statistical significance of regime-switching autoregressive models of measures of US economic activity. While the early literature did not find strong evidence for simple regime-switching models over the alternative of a constant parameter autoregression for US real GDP (e. g. [33]), later researchers have found stronger evidence using more complicated models of real GDP [57], alternative measures of economic activity [45], and multivariate techniques [63]. Examples of other studies finding statistical evidence in favor of regime-switching models include Garcia and Perron [34], who document regime switching in the conditional mean of an autoregression for the US real interest rate, and Guidolin and Timmermann [40], who find evidence of regime-switching in the conditional mean and volatility of UK equity returns.

This article surveys the literature surrounding regime-switching models, focusing primarily on Markov-switching models. The organization of the article is as follows. Section "Threshold and Markov-Switching Models of Regime Change"describes both threshold and Markov-switching models using a simple example. The article then focuses on Markov-switching models, with Sect. "Estimation of a Basic Markov-Switching Model" discussing estimation techniques for a basic model, Sect. "Extensions of the Basic Markov-Switching Model" surveying a number of primary extensions of the basic model, and Sect. "Specification Testing for Markov-Switching Models" surveying issues related to specification analysis. Section "Empirical Example: Identifying Business Cycle Turning Points" gives an empirical example, discussing how Markov-switching models can be used to identify turning points in the US business cycle. The article concludes by highlighting some particular avenues for future research.

Threshold and Markov-Switching Models of Regime Change

This section describes the threshold and Markov-switching approaches to modeling regime-switching using a specific example. In particular, suppose we are interested in modeling the sample path of a time series, $\{y_t\}_{t=1}^{T}$, where y_t is a scalar, stationary, random variable. A popular choice is an autoregressive (AR) model of order k:

$$y_t = \alpha + \sum_{j=1}^{k} \phi_j y_{t-j} + \varepsilon_t \,, \tag{1}$$

where the disturbance term, ε_t, is assumed to be normally distributed, so that $\varepsilon_t \sim N(0, \sigma^2)$. The AR($k$) model in (1) is a parsimonious description of the data, and has a long history as a tool for establishing stylized facts about the dynamic behavior of the time series, as well as an impressive record in forecasting.

In many cases however, we might be interested in whether the behavior of the time series changes across different periods of time, or regimes. In particular, we may be interested in the following regime-switching version of (1):

$$y_t = \alpha_{S_t} + \sum_{j=1}^{k} \phi_{j,S_t} y_{t-j} + \varepsilon_t \,, \tag{2}$$

where $\varepsilon_t \sim N(0, \sigma_{S_t}^2)$. In (2), the parameters of the AR(k) depend on the value of a discrete-valued state variable, $S_t = i, i = 1, \ldots, N$, which denotes the regime in operation at time t. Put simply, the parameters of the AR(k) model are allowed to vary among one of N different values over the sample period.

There are several items worth emphasizing about the model in (2). First, conditional on being inside of any particular regime, (2) is simply a constant parameter linear regression. Such models, which are commonly referred to as "piecewise linear", make up the vast majority of the applications of regime-switching models. Second, if the state variable were observed, the model in (2) is simply a linear regression model with dummy variables, a fact that will prove important in our discussion of how the parameters of (2) might be estimated. Third, although the specification in (2) allows for all parameters to switch across all regimes, more restrictive models are certainly possible, and indeed are common in applied work. For example, a popular model for time series of asset prices is one in which only the variance of the disturbance term is allowed to vary across regimes. Finally, the shifts in the parameters of (2) are modeled as occurring abruptly. An example of an alternative approach, in which parameter shifts are phased in gradually, can be found in the literature investigating "smooth transition" threshold models. Such models will not be described further here, but are discussed in detail in [93].

Threshold and Markov-switching models differ in the assumptions made about the state variable, S_t. Threshold models assume that S_t is a deterministic function of an observed variable. In most applications this variable is taken to be a particular lagged value of the process itself, in which case regime shifts are said to be "self-exciting". In particular, define $N-1$ "thresholds" as $\tau_1 < \tau_2 < \ldots < \tau_{N-1}$. Then, for a self-exciting threshold model, S_t is defined as follows:

$$\begin{aligned} S_t &= 1 && y_{t-d} < \tau_1 \,, \\ S_t &= 2 && \tau_1 \le y_{t-d} < \tau_2 \,, \\ &\ \vdots && \ \vdots \\ S_t &= N && \tau_{N-1} \le y_{t-d} \,. \end{aligned} \tag{3}$$

In (3), d is known as the "delay" parameter. In most cases S_t is unobserved by the econometrician, because the delay and thresholds, d and τ_i, are generally not observable. However, d and τ_i can be estimated along with other model parameters. [78] surveys classical and Bayesian approaches to estimation of the parameters of threshold models.

Markov-switching models also assume that S_t is unobserved. In contrast to threshold models however, S_t is assumed to follow a particular stochastic process, namely an N-state Markov chain. The evolution of Markov chains are described by their transition probabilities, given by:

$$\begin{aligned} &P(S_t = i | S_{t-1} = j, S_{t-2} = q, \ldots) \\ &\qquad = P(S_t = i | S_{t-1} = j) = p_{ij} \,, \end{aligned} \tag{4}$$

where, conditional on a value of j, we assume $\sum_{i=1}^{N} p_{ij} = 1$. That is, the process in (4) specifies a complete probability distribution for S_t. In the general case, the Markov process allows regimes to be visited in any order and for regimes to be visited more than once. However, restrictions can be placed on the p_{ij} to restrict the order of regime shifts. For example, [12] notes that the transition probabilities can be restricted in such a way so that the model in (2) becomes a "changepoint" model in which there are $N-1$ structural breaks in the model parameters. Finally, the vast majority of the applied literature has assumed that the transition probabilities in (4) evolve independently of lagged values of the series itself, so that

$$\begin{aligned} &P(S_t = i | S_{t-1} = j, S_{t-2} = q, \ldots, y_{t-1}, y_{t-2}, \ldots) \\ &\qquad = P(S_t = i | S_{t-1} = j) = p_{ij} \,, \end{aligned} \tag{5}$$

which is the polar opposite of the threshold process described in (3). For this reason, Markov-switching models are often described as having regimes that evolve "exogenously" of the series, while threshold models are said to have "endogenous" regimes. However, while popular in practice, the restriction in (5) is not necessary for estimation of the parameters of the Markov-switching model. Section "Extensions of the Basic Markov-Switching Model" of this article discusses models in which the transition probabilities of the Markov process are allowed to be partially determined by lagged values of the series.

The threshold and Markov-switching approaches are best viewed as complementary, with the "best" model likely to be application specific. Certain applications appear tailor-made for the threshold assumption. For example, we might have good reason to think that the behavior of time series such as an exchange rate or inflation will exhibit regime shifts when the series moves outside of certain thresholds, as this will trigger government intervention. The Markov-switching model might instead be the obvious choice when one does not wish to tie the regime shifts to the behavior of a particular observed variable, but instead wishes to let the data speak freely as to when regime shifts have occurred.

In the remainder of this article I will survey various aspects regarding the econometrics of Markov-switching models. For readers interested in learning more about threshold models, the survey article of Potter [78] is an excellent starting point.

Estimation of a Basic Markov-Switching Model

This section discusses estimation of the parameters of Markov-switching models. The existing literature has focused almost exclusively on likelihood-based methods for estimation. I retain this focus here, and discuss both maximum likelihood and Bayesian approaches to estimation. An alternative approach based on semi-parametric estimation is discussed in [4].

To aid understanding, we focus on a specific baseline case, which is the Markov-switching autoregression given in (2) and (5). We simplify further by allowing for $N = 2$ regimes, so that $S_t = 1$ or 2. It is worth noting that in many cases two regimes is a reasonable assumption. For example, in the literature using Markov-switching models to study business cycles phases, a two regime model, meant to capture an expansion and recession phase, is an obvious starting point that has been used extensively.

Estimation of Markov-switching models necessitates two additional restrictions over constant parameter models. First of all, the labeling of S_t is arbitrary, in that switching the vector of parameters associated with $S_t = 1$ and $S_t = 2$ will yield an identical model. A commonly used approach to normalize the model is to restrict the value of one of the parameters when $S_t = 1$ relative to its value when $S_t = 2$. For example, for the model in (2) we could restrict $\alpha_2 < \alpha_1$. For further details on the choice of normalization, see [49]. Second, the transition probabilities in (5) must be constrained to lie in [0,1]. One approach to implement this constraint, which will be useful in later discussion, is to use a probit specification for S_t. In particular, the value of S_t is assumed to be determined by the realization of a random variable, η_t, as follows:

$$S_t = \left\{ \begin{array}{lll} 1 & \text{if} & \eta_t < \gamma_{S_{t-1}} \\ 2 & \text{if} & \eta_t \geq \gamma_{S_{t-1}} \end{array} \right\} , \tag{6}$$

where $\eta_t \sim i.i.d.N(0,1)$. The specification in (6) depends on two parameters, γ_1 and γ_2, which determine the transition probabilities of the Markov process as follows:

$$\begin{array}{l} p_{1j} = P(\eta_t < \gamma_j) = \Phi(\gamma_j) \\ p_{2j} = 1 - p_{1j} \end{array} , \tag{7}$$

where $j = 1, 2$ and Φ is the standard normal cumulative distribution function.

There are two main items of interest on which to conduct statistical inference for Markov-switching models. The first are the parameters of the model, of which there are $2(k+3)$ for the two-regime Markov-switching autoregression. In the following we collect these parameters in the vector

$$\begin{aligned} \theta = (&\alpha_1, \phi_{1,1}, \phi_{2,1}, \ldots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \ldots, \\ &\phi_{k,2}, \sigma_2, \gamma_1, \gamma_2)' . \end{aligned} \tag{8}$$

The second item of interest is the regime indicator variable, S_t. In particular, as S_t is unobserved, we will be interested in constructing estimates of which regime was in operation at each point in time. These estimates will take the form of posterior probabilities that $S_t = i, i = 1, 2$. We assume that the econometrician has a sample of $T + k$ observations, $(y_T, y_{T-1}, y_{T-2}, \ldots, y_{-(k-1)})$. The series of observations available up to time t is denoted as $\Omega_t = (y_t, y_{t-1}, y_{t-2}, \ldots, y_{-(k-1)})$.

We begin with maximum likelihood estimation of θ. Maximum likelihood estimation techniques for various versions of Markov-switching regressions can be found in the existing literature of multiple disciplines, for example [52,76,79] in the speech recognition literature, and [16,41] in the econometrics literature. Here we focus on the presentation of the problem given in [41], who presents a simple iterative algorithm that can be used to construct the likelihood function of a Markov-switching autoregression, as well as compute posterior probabilities for S_t.

For a given value of θ, the conditional log likelihood function is given by:

$$L(\theta) = \sum_{t=1}^{T} \log f(y_t|\Omega_{t-1};\theta) \,. \tag{9}$$

Construction of the conditional log likelihood function then requires construction of the conditional density function, $f(y_t|\Omega_{t-1};\theta)$, for $t = 1,\ldots,T$. The "Hamilton Filter" computes these conditional densities recursively as follows: Suppose for the moment that we are given $P(S_{t-1} = j|\Omega_{t-1};\theta)$, which is the posterior probability that $S_{t-1} = j$ based on information observed through period $t-1$. Equations (10) and (11) can then be used to construct $f(y_t|\Omega_{t-1};\theta)$:

$$P(S_t = i|\Omega_{t-1};\theta) = \sum_{j=1}^{2} P\left(S_t = i|S_{t-1} = j, \Omega_{t-1};\theta\right) \cdot P\left(S_{t-1} = j|\Omega_{t-1};\theta\right) \,, \tag{10}$$

$$f\left(y_t|\Omega_{t-1};\theta\right) = \sum_{i=1}^{2} f\left(y_t|S_t = i, \Omega_{t-1};\theta\right) \cdot P\left(S_t = i|\Omega_{t-1};\theta\right) \,. \tag{11}$$

From (5), the first term in the summation in (10) is simply the transition probability, p_{ij}, which is known for any particular value of θ. The first term in (11) is the conditional density of y_t assuming that $S_t = i$, which, given the within-regime normality assumption for ε_t, is:

$$f(y_t|S_1 = i, \Omega_0;\theta) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{-\left(y_t - \alpha_i - \sum_{j=1}^{k} \phi_{j,i} y_{t-j}\right)^2}{2\sigma_i^2}\right) \,. \tag{12}$$

With $f(y_t|\Omega_{t-1};\theta)$ in hand, the next step is then to update (10) and (11) to compute $f(y_{t+1}|\Omega_t;\theta)$. To do so requires $P(S_t = i|\Omega_t;\theta)$ as an input, meaning we must update $P(S_t = i|\Omega_{t-1};\theta)$ to reflect the information contained in y_t. This updating is done using Bayes' rule:

$$P(S_t = i|\Omega_t;\theta) = \frac{f(y_t|S_t = i, \Omega_{t-1};\theta)P\left(S_t = i|\Omega_{t-1}\right)}{f\left(y_t|\Omega_{t-1};\theta\right)} \,, \tag{13}$$

where each of the three elements on the right-hand side of (13) are computable from the elements of (10) and (11). Given a value for $P(S_0 = i|\Omega_0;\theta)$ to initialize the filter, Eqs. (10) through (13) can then be iterated to construct $f(y_t|\Omega_{t-1};\theta)$, $t = 1,\ldots,T$, and therefore the log likelihood function, $L(\theta)$. The *maximum likelihood estimate* $\hat{\theta}_{\text{MLE}}$, is then the value of θ that maximizes $L(\theta)$, and can be obtained using standard numerical optimization techniques.

How do we set $P(S_0 = i|\Omega_0;\theta)$ to initialize the filter? As is discussed in [41], exact evaluation of this probability is rather involved. The usual practice, which is possible when S_t is an ergodic Markov chain, is to simply set $P(S_0 = i|\Omega_0;\theta)$ equal to the unconditional probability, $P(S_0 = i)$. For the two-regime case considered here, these unconditional probabilities are given by:

$$\begin{aligned} P(S_0 = 1) &= \frac{1-p_{22}}{2-p_{11}-p_{22}} \\ P(S_0 = 2) &= 1 - P(S_0 = 1) \,. \end{aligned} \tag{14}$$

Alternatively, $P(S_0 = i|\Omega_0;\theta)$ could be treated as an additional parameter to be estimated. See Hamilton [43] and Kim and Nelson [62] for further details.

An appealing feature of the Hamilton filter is that, in addition to the likelihood function, the procedure also directly evaluates $P(S_t = i|\Omega_t;\theta)$, which is commonly referred to as a "filtered" probability. Inference regarding the value of S_t is then sometimes based on $P(S_t = i|\Omega_t;\hat{\theta}_{\text{MLE}})$, which is obtained by running the Hamilton filter with $\theta = \hat{\theta}_{\text{MLE}}$. In many circumstances, we might also be interested in the so-called "smoothed" probability of a regime computed using all available data, or $P(S_t = i|\Omega_T;\theta)$. [54] presents an efficient recursive algorithm that can be applied to compute these smoothed probabilities.

We now turn to Bayesian estimation of Markov-switching models. In the Bayesian approach, the parameters θ are themselves assumed to be random variables, and the goal is to construct the posterior density for these parameters given the observed data, denoted $f(\theta|\Omega_T)$. In all but the simplest of models, this posterior density does not take the form of any well known density whose properties can be analyzed analytically. In this case, modern Bayesian inference usually proceeds by sampling the posterior density repeatedly to form estimates of posterior moments and other objects of interest. These estimates can be made arbitrarily accurate by increasing the number of samples taken from the posterior. In the case of Markov-switching models, Albert and Chib [1] demonstrate that samples from $f(\theta|\Omega_T)$ can be obtained using a simulation-based approach known as the Gibbs Sampler. The Gibbs Sampler, introduced by [37,38,89], is an algorithm that produces random samples from the joint density of a group of random variables by repeatedly sampling from the full set of conditional densities for the random variables.

We will sketch out the main ideas of the Gibbs Sampler in the context of the two-regime Markov-switching autoregression. It will prove useful to divide the parameter space into $\theta = (\theta_1', \theta_2')'$, where $\theta_1 = (\alpha_1, \phi_{1,1}, \phi_{2,1}, \ldots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \ldots, \phi_{k,2}, \sigma_2)'$ and $\theta_2 = (\gamma_1, \gamma_2)'$. Suppose it is feasible to simulate draws from the three conditional distributions, $f(\theta_1|\theta_2, \tilde{S}, \Omega_T)$, $f(\theta_2|\theta_1, \tilde{S}, \Omega_T)$, and $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$, where $\tilde{S} = (S_1, S_2, \ldots, S_T)'$. Then, conditional on arbitrary initial values, $\theta_2^{(0)}$ and $\tilde{S}^{(0)}$, we can obtain a draw of θ_1, denoted $\theta_1^{(1)}$, from $f(\theta_1|\theta_2^{(0)}, \tilde{S}^{(0)}, \Omega_T)$, a draw of θ_2, denoted $\theta_2^{(1)}$, from $f(\theta_2|\theta_1^{(1)}, \tilde{S}^{(0)}, \Omega_T)$, and a draw of $\tilde{S}$, denoted $\tilde{S}^{(1)}$, from $P(\tilde{S}|\theta_1^{(1)}, \theta_2^{(1)}, \Omega_T)$. This procedure can be iterated to obtain $\theta_1^{(j)}$, $\theta_2^{(j)}$, and $\tilde{S}^{(j)}$, for $j = 1, \ldots, J$. For large enough J, and assuming weak regularity conditions, these draws will converge to draws from $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$. Then, by taking a large number of such draws beyond J, one can estimate any feature of $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$, such as moments of interest, with an arbitrary degree of accuracy. For example, an estimate of $P(S_t = i|\Omega_T)$ can be obtained by computing the proportion of draws of $\tilde{S}$ for which $S_t = i$.

Why is the Gibbs Sampler useful for a Markov-switching model? It turns out that although $f(\theta|\Omega_t)$ and $P(\tilde{S}|\Omega_T)$ cannot be sampled directly, it is straightforward, assuming natural conjugate prior distributions, to obtain samples from $f(\theta_1|\theta_2, \tilde{S}, \Omega_T)$, $f(\theta_2|\theta_1, \tilde{S}, \Omega_T)$, and $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$. This is most easily seen for the case of θ_1, which, when $\tilde{S}$ is conditioning information, represents the parameters of a linear regression with dummy variables, a case for which techniques to sample the parameter posterior distribution are well established (Zellner 96). An algorithm for obtaining draws of $\tilde{S}$ from $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$ was first given in Albert and Chib [1], while Kim and Nelson [59] develop an alternative, efficient, algorithm based on the notion of "multi-move" Gibbs Sampling introduced in [6]. For further details regarding the implementation of the Gibbs Sampler in the context of Markov-switching models, see Kim and Nelson [62].

The Bayesian approach has a number of features that make it particularly attractive for estimation of Markov-switching models. First of all, the requirement of prior density functions for model parameters, considered by many to be a weakness of the Bayesian approach in general, is often an advantage for Bayesian analysis of Markov-switching models [42]. For example, priors can be used to push the model toward capturing one type of regime-switching vs. another. The value of this can be seen for Markov-switching models of the business cycle, for which the econometrician might wish to focus on portions of the likelihood surface related to business cycle switching, rather than those related to longer term regime shifts in productivity growth. Another advantage of the Bayesian approach is with regards to the inference drawn on S_t. In the maximum likelihood approach, the methods of [54] can be applied to obtain $P(S_t = i|\Omega_T; \hat{\theta}_{\text{MLE}})$. As these probabilities are conditioned on the maximum likelihood parameter estimates, uncertainty regarding the unknown values of the parameters has not been taken into account. By contrast, the Bayesian approach yields $P(S_t = i|\Omega_T)$, which is not conditional on a particular value of θ and thus incorporates uncertainty regarding the value of θ that generated the observed data.

Extensions of the Basic Markov-Switching Model

The basic, two-regime Markov-switching autoregression in (2) and (5) has been used extensively in the literature, and remains a popular specification in applied work. However, it has been extended in a number of directions in the substantial literature that follows [41]. This section surveys a number of these extensions.

The estimation techniques discussed in Sect. "Estimation of a Basic Markov-Switching Model" can be adapted in a straightforward manner to include several extensions to the basic Markov-switching model. For example, the filter used in (10) through (13) can be modified in obvious ways to incorporate the case of $N > 2$ regimes, as well as to allow y_t to be a vector of random variables, so that the model in (2) becomes a Markov-switching vector autoregression (MS-VAR). Hamilton [43] discusses both of these cases, while Krolzig [68] provides an extensive discussion of MS-VARs. [83] is a recent example of applied work using an MS-VAR with a large number of regimes. In addition, the (known) within-regime distribution of the disturbance term, ε_t, could be non-Gaussian, as in [23] or [45]. Further, the parameters of (2) could be extended to depend not just on S_t, but also on a finite number of lagged values of S_t, or even a second state variable possibly correlated with S_t. Indeed, such processes can generally be rewritten in terms of the current value of a single, suitably redefined, state variable. [58,66] provide examples of such a redefinition. For further discussion of all of these cases, see [43].

The specification for the transition probabilities in (5) restricted the probability $S_t = i$ to depend only on the value of S_{t-1}. However, in some applications we might think that these transition probabilities are driven in part by observed variables, such as the past evolution of the process. To this end, [21,28] develop Markov-switching models with time-varying transition probabilities (TVTP), in which the transition probabilities are allowed to vary depending on conditioning information. Suppose that z_t represents a vector of observed variables that are thought to influence the realization of the regime. The probit representation for the state process in (6) and (7) can then be extended as follows:

$$S_t = \left\{ \begin{array}{lll} 1 & \text{if} & \eta_t < \left(\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}\right) \\ 2 & \text{if} & \eta_t \geq \left(\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}\right) \end{array} \right\} , \tag{15}$$

with associated transition probabilities:

$$\begin{aligned} p_{1j}(z_t) &= P\big(\eta_t < (\gamma_j + z_t' \lambda_j)\big) = \Phi\big(\gamma_j + z_t' \lambda_j\big) \\ p_{2j}(z_t) &= 1 - p_{1j}(z_t) , \end{aligned} \tag{16}$$

where $j = 1,\ 2$ and Φ is again the standard normal cumulative distribution function. Estimation of the Markov-switching autoregression with TVTP is then straightforward. In particular, assuming that z_t contains lagged values of y_t or exogenous random variables, a maximum likelihood estimation proceeds by simply replacing p_{ij} with $p_{ij}(z_t)$ in the filter given in (10) through (13). Bayesian estimation of TVTP models via the Gibbs Sampler is also straightforward, and is discussed in [29]. Despite its intuitive appeal, the literature contains relatively few applications of the TVTP model. A notable example of the TVTP framework is found in Durland and McCurdy [24], Filardo and Gordon [29] and Kim and Nelson [59], who study business cycle "duration dependence", or whether the probability of a business cycle phase shift depends on how long the economy has been in the current phase. Other applications include Ang and Bekaert [2], who model regime-switches in interest rates, and Lo and Piger [69], who investigate sources of time-variation in the response of output to monetary policy actions.

The TVTP model is capable of relaxing the restriction that the state variable, S_t, is independent of the lagged values of the series, y_t, and thus of lagged values of the disturbance term, ε_t. Kim, Piger and Startz [65] consider a Markov-switching model in which S_t is also correlated with the contemporaneous value of ε_t, and is thus "endogenous". They model this endogenous switching by assuming that the shock to the probit process in (6), η_t, and ε_t are jointly normally distributed as follows:

$$\begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \sim N(0, \Sigma),\ \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} . \tag{17}$$

Kim, Piger and Startz [65] show that when $\rho \neq 0$, the conditional density in (12) is no longer Gaussian, but can be evaluated analytically. Thus, the likelihood function for the endogenous switching model can be evaluated with simple modifications to the recursive filter in (10) through (13). Tests of the null hypothesis that S_t is exogenous can also be implemented in a straightforward manner. Chib and Dueker [13] consider endogenous switching as in (17) from a Bayesian perspective.

The extensions listed above are primarily modifications to the stochastic process assumed to drive S_t. A more fundamental extension of (2) is to consider Markov-switching in time series models that are more complicated than simple autoregressions. An important example of this is a state-space model with Markov-switching parameters. Allowing for Markov-switching in the state-space representation for a time series is particularly interesting because a large number of popular time-series models can be given a state-space representation. Thus, incorporating Markov-switching into a general state-space representation immediately extends the Markov-switching framework to these models.

To aid discussion, consider the following Markov-switching state-space representation for a vector of R random variables, $Y_t = (y_{1t}, y_{2t}, \ldots, y_{Rt})'$, given as follows:

$$\begin{aligned} Y_t &= H_{S_t}' X_t + W_t \\ X_t &= A_{S_t} + F_{S_t} X_{t-1} + V_t \end{aligned} , \tag{18}$$

where $X_t = (x_{1t}, x_{2t}, \ldots, x_{Dt})'$, $W_t \sim N(0, B_{S_t})$ and $V_t \sim N(0, Q_{S_t})$. The parameters of the model undergo Markov switching, and are contained in the matrices $H_{S_t}, B_{S_t}, A_{S_t}, F_{S_t}, Q_{S_t}$. A case of primary interest is when some or all of the elements of X_t are unobserved. This is the case for a wide range of important models in practice, including models with moving average (MA) dynamics, unobserved components (UC) models, and dynamic factor models. However, in the presence of Markov-switching parameters, the fact that X_t is unobserved introduces substantial complications for construction of the likelihood function. In particular, as is discussed in detail in [54] and Kim and Nelson [62], exact construction of the conditional density $f(y_t|\Omega_{t-1};\theta)$ requires that one consider all possible permutations of the entire history of the state variable, $S_t, S_{t-1}, S_{t-2}, \ldots, S_1$. For even moderately sized values of t, this quickly becomes computationally infeasible.

To make inference via maximum likelihood estimation feasible, [54] develops a recursive filter that constructs an approximation to the likelihood function. This filter "collapses" the number of lagged regimes that are necessary to keep track of by approximating a nonlinear expectation with a linear projection. Kim and Nelson [62] provide a detailed description of the Kim [54] filter, as well as a number of examples of its practical use.

If one is willing to take a Bayesian approach to the problem, Kim and Nelson [59] show that inference can be conducted via the Gibbs Sampler without resorting to approximations. As before, the conditioning features of the Gibbs sampler greatly simplifies the analysis. For example, by conditioning on $\tilde{S} = (S_1, S_2, \ldots, S_T)'$, the model in (18) is simply a linear, Gaussian, state-space model with dummy variables, for which techniques to sample the posterior distribution of model parameters and the unobserved elements of X_t are well established [6]. Kim and Nelson [62] provide detailed descriptions of how the Gibbs Sampler can be implemented for a state-space model with Markov switching.

There are many applications of state space models with Markov switching. For example, a large literature uses UC models to decompose measures of economic output into trend and cyclical components, with the cyclical component often interpreted as a measure of the business cycle. Until recently, this literature focused on linear representations for the trend and cyclical components [14,51,72,94]. However, one might think that the processes used to describe the trend and cyclical components might display regime switching in a number of directions, such as that related to the phase of the business cycle or to longer-run structural breaks in productivity growth or volatility. A UC model with Markov switching in the trend and cyclical components can be cast as a Markov-switching state-space model as in (18). Applications of such regime-switching UC models can be found in [58,60,64,71,84]. Another primary example of a Markov-switching state-space model is a dynamic factor model with Markov-switching parameters, examples of which are given in [7,59]. Section"Empirical Example: Identifying Business Cycle Turning Points" presents a detailed empirical example of such a model.

Specification Testing for Markov-Switching Models

Our discussion so far has assumed that key elements in the specification of regime-switching models are known to the researcher. Chief among these is the number of regimes, N. However, in practice there is likely uncertainty about the appropriate number of regimes. This section discusses data-based techniques that can be used to select the value of N.

To fix ideas, consider a simple version of the Markov-switching model in (2):

$$y_t = \alpha_{S_t} + \varepsilon_t \,, \tag{19}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Consider the problem of trying to decide between a model with $N = 2$ regimes vs. the simpler model with $N = 1$ regimes. The model with one regime is a constant parameter model, and thus this problem can be interpreted as a decision between a model with regime-switching parameters vs. one without. An obvious choice for making this decision is to construct a test of the null hypothesis of $N = 1$ vs. the alternative of $N = 2$. For example, one might construct the likelihood ratio statistic:

$$LR = 2\left(L(\hat{\theta}_{\text{MLE}(2)}) - L(\hat{\theta}_{\text{MLE}(1)})\right) , \tag{20}$$

where $\hat{\theta}_{\text{MLE}(1)}$ and $\hat{\theta}_{\text{MLE}(2)}$ are the maximum likelihood estimates under the assumptions of $N = 1$ and $N = 2$ respectively. Under the null hypothesis there are three fewer parameters to estimate, α_2, γ_1 and γ_2, than under the alternative hypothesis. Then, to test the null hypothesis, one might be tempted to proceed by constructing a p-value for LR using the standard $\chi^2(3)$ distribution.

However, this final step is not justified, and can lead to very misleading results in practice. In particular, the standard conditions for LR to have an asymptotic χ^2 distribution include that all parameters are identified under the null hypothesis [17]. In the case of the model in (19), the parameters γ_1 and γ_2, which determine the transition probabilities p_{ij}, are not identified assuming the null hypothesis is true. In particular, if $\alpha_1 = \alpha_2$, then p_{ij} can take on any values without altering the likelihood function for the observed data. A similar problem exists when testing the general case of N vs. $N + 1$ regimes.

Fortunately, a number of contributions in recent years have produced asymptotically justified tests of the null hypothesis of N regimes vs. the alternative of $N + 1$ regimes. In particular, [33,50] provide techniques to compute asymptotically valid critical values for LR. Recently Carrasco, Hu and Ploberger [5] have developed an asymptotically optimal test for the null hypothesis of parameter constancy against the general alternative of Markov-switching parameters. Their test is particularly appealing because it does not require estimation of the model under the alternative hypothesis, as is the case with LR.

If one is willing to take a Bayesian approach, the comparison of models with N vs. $N + 1$ regimes creates no special considerations. In particular, one can proceed by computing standard Bayesian model comparison metrics, such as Bayes Factors or posterior odds ratios. Examples of such comparisons can be found in [11,63,78].

Empirical Example: Identifying Business Cycle Turning Points

This section presents an empirical example demonstrating how the Markov-switching framework can be used to model shifts between expansion and recession phases in the US business cycle. This example is of particular interest for two reasons. First, although Markov-switching models have been used to study a wide variety of topics, their most common application has been as formal statistical models of business cycle phase shifts. Second, the particular model we focus on here, a dynamic factor model with Markov-switching parameters, is of interest in its own right, with a number of potential applications.

The first presentation of a Markov-switching model of the business cycle is found in [41]. In particular, [41] showed that US real GDP growth could be characterized as an autoregressive model with a mean that switched between low and high growth regimes, where the estimated timing of the low growth regime corresponded closely to the dates of US recessions as established by the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER). This suggested that Markov-switching models could be used as tools to identify the timing of shifts between business cycle phases, and a great amount of subsequent analysis has been devoted toward refining and using the Markov-switching model for this task.

The model used in [41] was univariate, considering only real GDP. However, as is discussed in [22], a long emphasized feature of the business cycle is comovement, or the tendency for business cycle fluctuations to be observed simultaneously in a large number of economic sectors and indicators. This suggests that, by using information from many economic indicators, the identification of business cycle phase shifts might be sharpened. One appealing way of capturing comovement in a number of economic indicators is through the use of dynamic factor models, as popularized by [85,86]. However, these models assumed constant parameters, and thus do not model business cycle phase shifts explicitly.

To simultaneously capture comovement and business cycle phase shifts, [7] introduces Markov-switching parameters into the dynamic factor model of [85,86]. Specifically, defining $y^*_{rt} = y_{rt} - \bar{y}_r$ as the demeaned growth rate of the rth economic indicator, the dynamic factor Markov-switching (DFMS) model has the form:

$$y^*_{rt} = \beta_r c_t + e_{rt} \,. \tag{21}$$

In (21), the demeaned first difference of each series is made up of a component common to each series, given by the dynamic factor c_t, and a component idiosyncratic to each series, given by e_{rt}. The common component is assumed to follow a stationary autoregressive process:

$$\phi(L)(c_t - \mu_{S_t}) = \varepsilon_t \,, \tag{22}$$

where $\varepsilon_t \sim i.i.d.N(0,1)$. The unit variance for ε_t is imposed to identify the parameters of the model, as the factor loading coefficients, β_r, and the variance of ε_t are not separately identified. The lag polynomial $\phi(L)$ is assumed to have all roots outside of the unit circle. Regime switching is introduced by allowing the common component to have a Markov-switching mean, given by μ_{S_t}, where $S_t = \{1,2\}$. The regime is normalized by setting $\mu_2 < \mu_1$. Finally, each idiosyncratic component is assumed to follow a stationary autoregressive process:

$$\theta_r(L)e_{rt} = \omega_{rt} \,. \tag{23}$$

where $\theta_r(L)$ is a lag polynomial with all roots outside the unit circle and $\omega_{rt} \sim N(0, \sigma^2_{\omega,r})$.

[7] estimates the DFMS model for US monthly data on non-farm payroll employment, industrial production, real manufacturing and trade sales, and real personal income excluding transfer payments, which are the four monthly variables highlighted by the NBER in their analysis of business cycles. The DFMS model can be cast as a state-space model with Markov switching of the type discussed in Sect. "Extensions of the Basic Markov-Switching Model". Chauvet estimates the parameters of the model via maximum likelihood, using the approximation to the likelihood function given in [54]. Kim and Nelson [59] instead use Bayesian estimation via the Gibbs Sampler to estimate the DFMS model.

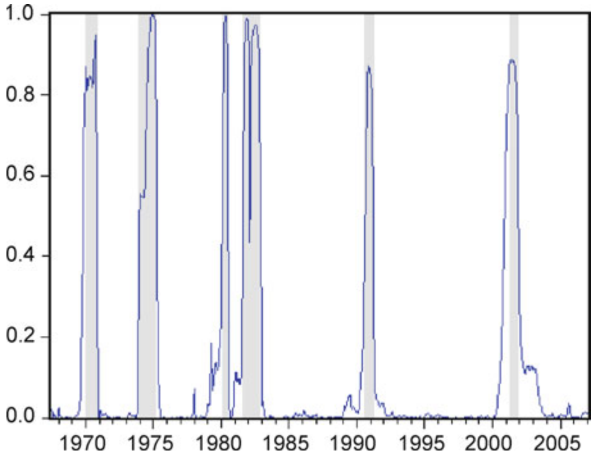
Here I update the estimation of the DFMS model presented in [59] to a sample period extending from February 1967 through February 2007. For estimation, I use the Bayesian Gibbs Sampling approach, with prior distributions and specification details identical to those given in [59]. Figure 1 displays $P(S_t = 2|\Psi_T)$ obtained from the Gibbs Sampler, which is the estimated probability that the low growth regime is active. For comparison, Fig. 1 also indicates NBER recession dates with shading.

There are two items of particular interest in Fig. 1. First of all, the estimated probability of the low growth regime is very clearly defined, with $P(S_t = 2|\Psi_T)$ generally close to either zero or one. Indeed, of the 481 months in the sample, only 32 had $P(S_t = 2|\Psi_T)$ fall between 0.2 and 0.8. Second, $P(S_t = 2|\Psi_T)$ is very closely aligned with NBER expansion and recession dates. In particular, $P(S_t = 2|\Psi_T)$ tends to be very low during NBER expansion phases and very high during NBER recession phases.

Figure 1 demonstrates the added value of employing the DFMS model, which considers the comovement between multiple economic indicators, over models considering only a single measure of economic activity. In particular, results for the Markov-switching autoregressive model of real GDP presented in [41] were based on a data

Econometrics: Models of Regime Changes, Table 1
Dates of Business Cycle Turning Points Produced by NBER and Dynamic Factor Markov-Switching Model

Peaks			Troughs		
DFMS	NBER	Discrepancy	DFMS	NBER	Discrepancy
Oct 1969	Dec 1969	2M	Nov 1970	Nov 1970	0M
Dec 1973	Nov 1973	−1M	Mar 1975	Mar 1975	0M
Jan 1980	Jan 1980	0M	Jun 1980	Jul 1980	1M
Jul 1981	Jul 1981	0M	Nov 1982	Nov 1982	0M
Aug 1990	Jul 1990	−1M	Mar 1991	Mar 1991	0M
Nov 2000	Mar 2001	4M	Nov 2001	Nov 2001	0M



Econometrics: Models of Regime Changes, Figure 1
Probability of US Recession from Dynamic Factor Markov-Switching Model

sample ending in 1984, and it is well documented that Hamilton's original model does not perform well for capturing the two NBER recessions since 1984. Subsequent research has found that allowing for structural change in the residual variance parameter [61,70] or omitting all linear dynamics in the model [1,9] improves the Hamilton model's performance. By contrast, the results presented here suggest that the DFMS model accurately identifies the NBER recession dates without a need for structural breaks or the omission of linear dynamics.

In some cases, we might be interested in converting $P(S_t = 2|\Psi_T)$ into a specific set of dates establishing the timing of shifts between business cycle phases. To do so requires a rule for establishing whether a particular month was an expansion month or a recession month. Here we consider a simple rule, which categorizes any particular month as an expansion month if $P(S_t = 2|\Psi_T) \leq 0.5$ and a recession month if $P(S_t = 2|\Psi_T) > 0.5$. Table 1 displays the dates of turning points between expansion and recession phases (business cycle peaks), and the dates of turning points between recession and expansion phases (business cycle troughs) that are established by this rule. For comparison, Table 1 also lists the NBER peak and trough dates.

Table 1 demonstrates that the simple rule applied to $P(S_t = 2|\Psi_T)$ does a very good job of matching the NBER peak and trough dates. Of the twelve turning points in the sample, the DFMS model establishes eleven within two months of the NBER date. The exception is the peak of the 2001 recession, for which the peak date from the DFMS model is four months prior to that established by the NBER. In comparing peak and trough dates, the DFMS model appears to do especially well at matching NBER trough dates, for which the date established by the DFMS model matches the NBER date exactly in five of six cases.

Why has the ability of Markov-switching models to identify business cycle turning points generated so much attention? There are at least four reasons. First, it is sometimes argued that recession and expansion phases may not be of any intrinsic interest, as they need not reflect any real differences in the economy's structure. In particular, as noted by [95], simulated data from simple, constant parameter, time-series models, for which the notion of separate regimes is meaningless, will contain episodes that look to the eye like"recession" and "expansion" phases. By capturing the notion of a business cycle phase formally inside of a statistical model, the Markov-switching model is then able to provide statistical evidence as to the extent to which business cycle phases are a meaningful concept. Second, although the dates of business cycle phases and their associated turning points are of interest to many economic researchers, they are not compiled in a systematic fashion for many economies. Markov-switching models could then be applied to obtain business cycle turning point dates for these economies. An example of this is given in [74], who use Markov-switching models to establish business cycle phase dates for US states. Third, if economic time-series do display different behavior over business cycle phases, then Markov-switching models designed to capture such differences might be exploited to obtain more accurate forecasts of economic activity. Finally, the current probability of a new economic turning point is likely of substantial interest to economic policymakers. To this end, Markov-switching models can be used for "real-time" monitoring of new business cycle phase shifts. Indeed, Chauvet and Piger [10] provide evidence that Markov-switching models are often quicker to establish US business cycle turning points, particularly at business cycle troughs, than is the NBER. For additional analysis of the ability of regime-switching models to establish turning points in real time, see [8,9].

Future Directions

Research investigating applied and theoretical aspects of regime-switching models should be an important component of the future research agenda in macroeconomics and econometrics. In this section I highlight three directions for future research which are of particular interest.

To begin, additional research oriented toward improving the forecasting ability of regime-switching models is needed. In particular, given that regime-switching models of economic data contain important deviations from traditional, constant parameter, alternatives, we might expect that they could also provide improved out-of-sample forecasts. However, as surveyed in [15], the forecasting improvements generated by regime-switching models over simpler alternatives is spotty at best. That this is true is perhaps not completely surprising. For example, the ability of a Markov-switching model to identify regime shifts in past data does not guarantee that the model will do well at detecting regime shifts quickly enough in real time to generate improved forecasts. This is particularly problematic when regimes are short lived. Successful efforts to improve the forecasting ability of Markov-switching models are likely to come in the form of multivariate models, which can utilize additional information for quickly identifying regime shifts.

A second potentially important direction for future research is the extension of the Markov-switching dynamic factor model discussed in Sects."Extensions of the Basic Markov-Switching Model" and "Empirical Example: Identifying Business Cycle Turning Points" to settings with a large cross-section of data series. Indeed, applications of the DFMS model have been largely restricted to a relatively small number of variables, such as in the model of the US business cycle considered in Sect. "Empirical Example: Identifying Business Cycle Turning Points". However, in recent years there have been substantial developments in the analysis of dynamic factor models comprising a large number of variables, as in [31,32,87,88,92]. Research extending the regime-switching framework to such "big data" factor models will be of substantial interest.

Finally, much remains to be done incorporating regime-switching behavior into structural macroeconomic models. A number of recent studies have begun this synthesis by considering the implications of regime-switches in the behavior of a fiscal or monetary policymaker for the dynamics and equilibrium behavior of model economies [18,19,20,26,27]. This literature has already yielded a number of new and interesting results, and is likely to continue to do so as it expands. Less attention has been paid to reconciling structural models with a list of new "stylized facts" generated by the application of regime-switching models in reduced-form settings. As one example, there is now a substantial list of studies, including [3,45,57,58,82], and Kim and Nelson [60] finding evidence that the persistence of shocks to key macroeconomic variables varies dramatically over business cycle phases. However, such an asymmetry is absent from most modern structural macroeconomic models, which generally possess a symmetric propagation structure for shocks. Research designed to incorporate and explain business cycle asymmetries and other types of regime-switching behavior inside of structural macroeconomic models will be particularly welcome.

Bibliography

1. Albert J, Chib S (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J Bus Econ Stat 11:1–15
2. Ang A, Bekaert G (2002) Regime switches in interest rates. J Bus Econ Stat 20:163–182
3. Beaudry P, Koop G (1993) Do recessions permanently change output? J Monet Econ 31:149–163
4. Campbell SD (2002) Specification testing and semiparametric estimation of regime switching models: An examination of the us short term interest rate. Brown University Department of Economics Working Paper #2002–26, Providence
5. Carrasco M, Hu L, Ploberger W (2004) Optimal test for Markov switching. Working paper, University of Rochester, Rochester
6. Carter CK, Kohn P (1994) On Gibbs sampling for state space models. Biometrica 81:541–553
7. Chauvet M (1998) An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switching. Int Econ Rev 39:969–996
8. Chauvet M, Hamilton J (2006) Dating Business Cycle Turning Points. In: Milas C, Rothman P, van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, North Holland
9. Chauvet M, Piger J (2003) Identifying Business Cycle Turning Points in Real Time. Fed Res Bank St. Louis Rev 85:47–61
10. Chauvet M, Piger J (2004) A Comparison of the Real-Time Performance of Business Cycle Dating Methods. J Bus Econ Stat 26:42–49
11. Chib S (1995) Marginal Likelihood from the Gibbs Output. J Am Stat Assoc 90:1313–1321

12. Chib S (1998) Estimation and Comparison of Multiple Change-Point Models. J Econ 86:221–241
13. Chib S, Dueker M (2004) Non-Markovian Regime Switching with Endogenous States and Time Varying State Strengths. Federal Reserve Bank of St. Louis Working Paper #2004–030A, St. Louis
14. Clark PK (1987) The Cyclical Component of US Economic Activity. Quart J Econ 102:797–814
15. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-Linear Models. Int J Forecast 20:169–183
16. Cosslett SR, Lee LF (1985) Serial Correlation in Discrete Variable Models. J Econ 27:79–97
17. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254
18. Davig T, Leeper E (2005) Generalizing the Taylor Principle. Am Econ Rev 97:607–635
19. Davig T, Leeper E (2006) Endogenous Monetary Policy Regime Change. In: Reichlin L, West KD (eds) International Seminar on Macroeconomics. MIT Press, Cambridge
20. Davig T, Leeper E, Chung H (2004) Monetary and Fiscal Policy Switching. J Mon Credit Bank 39:809–842
21. Diebold FX, Lee JH, Weinbach G (1994) Regime Switching with Time-Varying Transition Probabilities. In: Hargreaves C (ed) Non-stationary Time Series Analysis and Cointegration. Oxford University Press, Oxford UK
22. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. Rev Econ Stat 78:67–77
23. Dueker M (1997) Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility. J Bus Econ Stat 15:26–34
24. Durland JM, McCurdy TH (1994) Duration Dependent Transitions in a Markov Model of USGNP Growth. J Bus Econ Stat 12:279–288
25. Evans M, Wachtel P (1993) Inflation Regimes and the Sources of Inflation Uncertainty. J Mon Credit Bank 25:475–511
26. Farmer REA, Waggoner DF, Zha T (2006) Indeterminacy in a Forward Looking Regime Switching Model. NBER working paper no. 12540, Cambridge
27. Farmer REA, Waggoner DF, Zha T (2007) Understanding the New-Keynesian Model when Monetary Policy Switches Regimes. NBER working paper no. 12965, Cambridge
28. Filardo AJ (1994) Business-Cycle Phases and Their Transitional Dynamics. J Bus Econ Stat 12:299–308
29. Filardo AJ, Gordon SF (1998) Business Cycle Durations. J Econ 85:99–123
30. Forni M, Hallin M, Lippi F, Reichlin L (2000) The Generalized Dynamic Factor Model: Identification and Estimation. Rev Econ Stat 82:540–554
31. Forni M, Hallin M, Lippi F, Reichlin L (2002) The generalized dynamic factor model: consistency and convergence rates. J Econ 82:540–554
32. Forni M, Hallin M, Lippi F, Reichlin L (2005) The generalized dynamic factor model: one-sided estimation and forecasting. J Am Stat Assoc 100:830–840
33. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models, Int Econ Rev 39:763–788
34. Garcia R, Perron P (1996) An Analysis of the Real Interest Rate under Regime Shifts. Rev Econ Stat 78:111–125
35. Garcia R, Schaller H (2002) Are the Effects of Monetary Policy Asymmetric? Econ Inq 40:102–119
36. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford
37. Gelfand AE, Smith AFM (1990) Sampling-Based Approaches to Calculating Marginal Densities. J Am Stat Assoc 85:398–409
38. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Trans Patt Anal Machine Int 6:721–741
39. Goldfeld SM, Quandt RE (1973) A Markov Model for Switching Regressions. J Econ 1:3–16
40. Guidolin M, Timmermann A (2005) Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns. Econ J 115:111–143
41. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57:357–384
42. Hamilton JD (1991) A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions. J Bus Econ Statistics 9:27–39
43. Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton NJ
44. Hamilton JD (2005) Regime-Switching Models. In: Durlauf S, Blume L (eds) New Palgrave Dictionary of Economics, 2nd edn. Palgrave McMillan Ltd, Hampshire
45. Hamilton JD (2005) What's Real About the Business Cycle? Fed Res Bank St. Louis Rev 87:435–452
46. Hamilton JD, Lin G (1996) Stock Market Volatility and the Business Cycle. J Appl Econ 11:573–593
47. Hamilton JD, Raj B (2002) New Directions in Business Cycle Research and Financial Analysis. Empir Econ 27:149–162
48. Hamilton JD, Susmel R (1994) Autoregressive Conditional Heteroskedasticity and Changes in Regime. J Econ 64:307–333
49. Hamilton, Waggoner JD DF, Zha T (2004) Normalization in Econometrics. Econ Rev 26:221–252
50. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. J Appl Econ 7:S61–S82
51. Harvey AC (1985) Trends and Cycles in Macroeconomic Time Series. J Bus Econ Stat 3:216–227
52. Juang BH, Rabiner LR (1985) Mixture Autoregressive Hidden Markov Models for Speech Signals. IEEE Trans Acoust Speech Signal Proc ASSP-30:1404–1413
53. Kaufmann S (2002) Is there an Asymmetric Effect of Monetary Policy Over Time? A Bayesian Analysis using Austrian Data. Empir Econ 27:277–297
54. Kim CJ (1994) Dynamic linear models with Markov-switching. J Econ 60:1–22
55. Kim CJ (2004) Markov-Switching Models with Endogenous Explanatory Variables. J Econ 122:127–136
56. Kim CJ, Morley J, Nelson C (2002) Is there a Positive Relationship between Stock Market Volatility and the Equity Premium? J Money Cred Bank 36:339–360
57. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the Permanent Effects of Recessions. J Appl Econ 20:291–309
58. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. Empir Econ 27:163–183
59. Kim CJ, Nelson CR (1998) Business Cycle Turning Points, a New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime Switching. Rev Econ Stat 80:188–201

60. Kim CJ, Nelson CR (1999b) Friedman's Plucking Model of Business Fluctuations: Tests and Estimates of Permanent and Transitory Components. J Money Cred Bank 31:317–34
61. Kim CJ, Nelson CR (1999c) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econ Stat 81:608–616
62. Kim CJ, Nelson C (1999a) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
63. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. Int Econ Rev 42:989–1013
64. Kim CJ, Piger J (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1189–1211
65. Kim CJ, Piger J, Startz R (2003) Estimation of Markov Regime-Switching Regression Models with Endogenous Switching. J Econom, (in press)
66. Kim CJ, Piger J, Startz R (2007) The Dynamic Relationship Between Permanent and Transitory Components of US Business Cycles. J Money Cred Bank 39:187–204
67. Koop G, Potter SM (1999) Bayes Factors and Nonlinearity: Evidence from Economic Time Series. J Econ 88:251–281
68. Krolzig HM (1997) Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin
69. Lo M, Piger J (2005) Is the Response of Output to Monetary Policy Asymmetric? Evidence from a Regime-Switching Coefficients Model. J Money Cred Bank 37:865–887
70. McConnell MM, Quiros GP (2000) Output Fluctuations in the United States: What has Changed Since the Early (1980s)? Am Econ Rev 90:1464–1476
71. Mills TC, Wang P (2002) Plucking Models of Business Cycle Fluctuations: Evidence from the G-7 Countries. Empir Econ 27:255–276
72. Morley JC, Nelson CR, Zivot E (2003) Why Are the Beveridge-Nelson and Unobserved-Components Decompositions of GDP So Different? Review Econ Stat 85:235–243
73. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. J Appl Econ 15:27–44
74. Owyang MT, Piger J, Wall HJ (2005) Business Cycle Phases in US States. Rev Econ Stat 87:604–616
75. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
76. Poritz AB (1982) Linear Predictive Hidden Markov Models and the Speech Signal. Acoustics, Speech and Signal Processing IEEE Conference on ICASSP '82, vol 7:1291–1294
77. Potter SM (1995) A Nonlinear Approach to US GNP. J Appl Econ 10:109–125
78. Potter SM (1999) Nonlinear Time Series Modelling: An Introduction. J Econ Surv 13:505–528
79. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc IEEE 77:257–286
80. Rapach DE, Wohar ME (2002) Regime Changes in International Real Interest Rates: Are They a Monetary Phenomenon? J Mon Cred Bank 37:887–906
81. Ravn M, Sola M (2004) Asymmetric Effects of Monetary Policy in the United States. Fed Res Bank St. Louis Rev 86:41–60
82. Sichel DE (1994) Inventories and the three phases of the business cycle. J Bus Econ Stat 12:269–277
83. Sims C, Zha T (2006) Were there Regime Changes in US Monetary Policy? Am Econ Rev 96:54–81
84. Sinclair T (2007) Asymmetry in the Business Cycle: A New Unobserved Components Model. George Washington University working paper, Washington
85. Stock JH, Watson MW (1989) New Indexes of Coincident and Leading Economic Indicators. NBER Macroeconomics Annual 4:351–393
86. Stock JH, Watson MW (1991) A Probability Model of the Coincident Economic Indicators. In: Lahiri K, Moore GH (eds) Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge University Press, Cambridge
87. Stock JH, Watson MW (2002a) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
88. Stock JH, Watson MW (2002b) Macroeconomic Forecasting Using Diffusion Indexes (with James H Stock). J Bus Econ Stat 20(2):147–162
89. Tanner M, Wong W (1987) The Calculation of Posterior Distributions by Data Augmentation. J Am Stat Assoc 82:528–550
90. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time-series analysis. J Forecast 13:109–131
91. Tong H (1983) Threshold models in non-linear time series analysis, Lecture Notes in Statistics, No. 21. Springer, Heidelberg
92. Turner CM, Startz R, Nelson CR (1989) A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market. J Financ Econ 25:3–22
93. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroecon Dyn 3:311–340
94. Watson MW (1986) Univariate Detrending Methods with Stochastic Trends. J Monet Econ 18:49–75
95. Watson MW (2005) Comment on James D Hamilton's 'What's Real about the Business Cycle?' Fed Res Bank St. Louis Rev 87:453–458
96. Zellner A (1971) An Introduction to Bayesian Inference in Econometrics. Wiley, New York

Econometrics: Non-linear Cointegration

JUAN-CARLOS ESCANCIANO[1], ALVARO ESCRIBANO[2]
[1] Department of Economics, Indiana University, Bloomington, USA
[2] Department of Economics, Universidad Carlos III de Madrid, Madrid, Spain

Article Outline

Glossary

Cointegration Cointegration is an econometric property relating time series variables. If two or more series are themselves nonstationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

Short memory A time series is said to be short memory if its information decays through time. In particular, we say that a variable is short memory in mean (in distribution), if the conditional mean (distribution) of the variable at time t given the information at time $t-h$ converges to a constant (to an unconditional distribution) as h diverges to infinity. Shocks in short memory time series have transitory effects.

Extended memory A time series is said to be extended memory in mean (in distribution), if it is not short memory in mean (distribution). Shocks in extended memory time series have permanent effects.

Nonlinear cointegration If two or more series are of extended memory, but a nonlinear transformation of them is short memory, then the series are said to be nonlinearly cointegrated.

Error correction model An Error Correction Model is a dynamic model in which the rate of growth of the variables in any period is related to the previous period's gap from long-run equilibrium.

Definition of the Subject

This paper is a selective review of the literature on nonlinear cointegration and nonlinear error correction models. The concept of cointegration plays a major role in macroeconomics, finance and econometrics. It was introduced by Granger in [42] and since then, it has achieved immense popularity among econometricians and applied economists. In fact in 2003 the Royal Swedish Academy of Science gave the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel to C. W. J. Granger for his contribution to the analysis of economic relationships based on cointegrated variables. In this paper we discuss the nonlinear extensions of the linear cointegration theory. Some authors consider nonlinear cointegration as a particular case of nonlinear error correction models. Although both concepts are related, we believe that it is useful to distinguish between them. After making this point clear, by relating linear and nonlinear error correction models, we discuss alternative measures of temporal dependence (memory) and co-dependence that are useful to characterize the usual notion of integration of order zero, $I(0)$, and cointegration in nonlinear contexts. We discuss parametric and nonparametric notions of nonlinear cointegration. Finally, we conclude pointing out several lines of research that we think are promising in nonlinear and nonstationary contexts and therefore deserve further analysis.

Introduction

Granger in [42] introduced the concept of *cointegration* in a linear context; for further development see [20,64,65,85]. The alternative ways to deal with integrated and cointegrated series are now clear only in the linear context; see for example [43,52,57,59,67,77,105].

In macroeconomic and financial applications there are many cases where *nonlinearities* have been found in nonstationary contexts and therefore, there is a need for a theoretical justification of those empirical results. To reach this goal is not an easy target since the usual difficulties analyzing nonlinear time series models within a stationary and ergodic framework are enhanced in nonstationary contexts.

The purpose of this survey on *nonlinear cointegration* is to give a selected overview on the state of the art of econometrics that simultaneously analyzes nonstationarities and nonlinearities. The structure of this paper is the following: Sect. "Linear Measures of Memory and Linear Error Correction Models" discusses linear concepts of memory and dependence, cointegrated and error correction models. Section "Nonlinear Error Correction (NEC) Models" introduces nonlinear error correction models. Section "Nonlinear Cointegration" investigates nonlinear measures of memory and dependence and nonlinear cointegration. Finally, Sect. "Future Directions" concludes and mentions some open questions for future research.

Linear Measures of Memory and Linear Error Correction Models

The time series x_t is *integrated of order d*, denoted $x_t \sim I(d)$, if $\Delta^d x_t = (1-L)^d x_t \sim I(0)$, where L is the lag operator such that $L^k x_t = x_{t-k}$ and d is an integer number. Here $I(0)$ denotes a covariance stationary short memory process with positive and bounded spectral density. We can extrapolate the concepts of integration to the *fractional case* where now d is not an integer but a real number. However, in this paper we will not cover fractional integration nor fractional cointegration; see Chapter 9.4.1 in [103] for a review.

Following the ideas of the seminal paper of [42], the most simple *definition of cointegration* could be the following; we say that two $I(1)$ series, y_t and x_t, are *cointegrated* if there is a linear combination $(1, -\beta)(y_t, x_t)'$ that is $I(0)$; $z_t = y_t - \beta x_t$ is $I(0)$, but any other linear combination, $z_t = y_t - \alpha x_t$, is $I(1)$ where $\alpha \neq \beta$. For simplicity, through all this paper we will assume a bivariate system with a single cointegrating vector. Notice that the second condition of the above definition of cointegration ($z_t = y_t - \alpha x_t$, is $I(1)$ for any $\alpha \neq \beta$) is redundant in the linear context. However, it will be useful for the identification of the cointegrating vector in nonlinear cointegration, see Definition 3.

A nonparametric characterization of cointegration was introduced by [1] and [78]. Let $x_t,\ y_t$ be the two $I(d)$ time series of interest, $d = 1$, and let $\gamma_{yx}(\tau, t)$ represent the *cross-covariance function* (*CCF*) of $x_t,\ y_t$, defined by $\gamma_{yx}(\tau, t) = \operatorname{cov}\left(y_t, x_{t-\tau}\right)$, where we make explicit the time dependence in $\gamma_{yx}(\tau, t)$ to allow for some degree of heterogeneity in the series. Similarly, define $\gamma_x(\tau, t) = \operatorname{cov}(x_t, x_{t-\tau})$. Cointegration implies that the rates of convergence of $\gamma_{yx}(\tau, t)$ and $\gamma_x(\tau, t)$ should be the same as τ increases without bound and $\tau = o(t)$. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in terms of the long-run linear forecast of y_t. For example, suppose $x_t, y_t \sim I(1)$ and $z_t = y_t - \beta x_t$ a sequence of independent and identically distributed (i.i.d.) random variables independent of x_t. In this case, $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = \beta$ for all $\tau, t, \tau \leqslant t$. In more general cases, z_t might have serial correlation and might not be independent of x_t and therefore, the constancy of this ratio will only take place for τ's beyond some value, see the Monte Carlo simulation results in [78]. On the other hand, in the spurious cointegration case where $x_t,\ y_t$ are stochastically independent, $\lim_{\tau\to\infty} \gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = 0$ for all $\tau, t, \tau \leqslant t$, therefore the ratio $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t)$ is consistent against this type of spurious alternative hypothesis. As we will see later on, this notion of cointegration accepts nonlinear generalizations (nonlinear cointegration).

The most simple version of *Granger's Representation Theorem*, see [20], states that two series y_t and x_t are cointegrated if and only if they have an error correction representation, see Eqs (1a)–(1c) below. Therefore, either x_t Granger-causes y_t or y_t Granger-causes x_t or both.

Let y_t and x_t be two $I(1)$ series, where x_t is a pure random walk and y_t is generated by the following linear error correction (EC) model (1a) with linear cointegration (1b),

$$\Delta y_t = \psi_1 \Delta x_t + \gamma z_{t-1} + \upsilon_t \tag{1a}$$

$$y_t = \beta x_t + z_t \tag{1b}$$

$$\Delta x_t = \varepsilon_t \tag{1c}$$

where all the random error terms $(\upsilon_t, z_t, \varepsilon_t)$ are $I(0)$. The errors z_t of (1b) form the error correction terms of (1a) and have usually more persistence (longer memory) than the other two random error terms $(\upsilon_t, \varepsilon_t)$. Therefore, in the system of Eqs. (1a)–(1c), x_t Granger-causes y_t but not the other way around.

Notice that we can write Eq. (1a), with $\phi_1 = \psi_1 - \beta$, in an equivalent way that will be very useful to introduce later on nonlinear error correction (NEC) models,

$$z_t = z_{t-1} + \phi_1 \Delta x_t + \gamma z_{t-1} + \upsilon_t\,. \tag{2}$$

Several alternative estimation procedures have been discussed in the literature to estimate the cointegrating parameter β:

i) Maximum likelihood approach of [66] and [7]. Assumes that the conditional distribution of y given x and the lagged values of x and y is *Normal* and that the bivariate data generating process (DGP) of y and x is a *VAR* of *finite autoregressive order k*, VAR(k) in error correction form. Furthermore, if the contemporaneous x-variable is *weakly exogenous*, then the partial maximum likelihood estimators is obtained by nonlinear least squares (NLS) on the error correction model obtained substituting (1b) in (1a). [98] and [38], derived the asymptotic properties of the NLS estimator of the error correction model (1a) and (1b), without the Normality assumption.

ii) *Two-step approach* of Engle and Granger, see [20]. In the first step, Eq. (1b) is estimated by ordinary least squares (OLS) to get the residuals (z). In the second step, Eq. (1a) is estimated by OLS after substituting z_{t-1} by the corresponding lagged residuals from the first step. The OLS estimator of the first step is *super-consistent but biased*, and the limiting distribution depends on nuisance parameters. However, if z_t is serially uncorrelated and x_t is strictly exogenous, then the OLS estimator in (1b) coincides with the fully modified estimator and therefore it is asymptotically efficient, see [38].

iii) *Fully modified OLS*, FM-OLS. This is a 2-step procedure developed by [79,80,86], and [81]. In the *first step*, Eq. (1b) is estimated by OLS. In the second step, semiparametric corrections are made for the *serial correlation* of the residuals z_t and for the *endogeneity* of the x-regressors. Under general conditions the fully modified estimator, is *asymptotically efficient*. The small sample behavior of these estimators was analyzed by Monte Carlo simulations by [53,56,57,58,71,86].

iv) *Fully modified instrumental variable* estimator, FM-IV, of [71,86]. [78] showed that their nonparametric notion of cointegration has an instrumental variable (IV) interpretation if the instruments are the lagged values of x. Furthermore, they showed that choosing those instruments has an extra advantage; we do not need to make the usual two corrections (endogeneity and serial correlation) to obtain a fully modified estimator. This particular IV-estimator has important advantages (bias reductions) over OLS in small samples.

v) Recently [89] also studied the asymptotic properties of instrumental variables estimators (IV) in a fractional cointegration context, as in [78]. They propose to use IV estimates based on single equations estimation like (1b) employing exclusion and normalization restrictions, without correcting for the serial correlation of z_t.

vi) [100] and [91], suggested a parametric correction for the endogeneity of the regressor (x_t) when estimating (1b) by OLS. The idea, based on the work of [97] about testing for causality, is to include additional future and past values of the Δx_t in Eq. (1b) when estimating it by OLS.

vii) [87] proposed to add integral error correction terms (lagged values of the EC terms), to the procedure described in *vi)* in order to parsimoniously correct for serial correlation.

[63], using Monte Carlo simulations, compares some of these parametric and semi parametric estimators of the cointegrating vector. In the context of normally distributed errors, [63] recommends to model explicitly the dynamics instead of using nonparametric corrections based on fully modified estimators.

Nonlinear Error Correction (NEC) Models

There are interesting macroeconomic applications where nonlinearities have been found in nonstationary contexts. The first example of a nonlinear error correction (NEC) model is the UK money demand from 1878 to 1970 of [25,27]. Later on [61] used this nonlinear error correction strategy in their money demand estimation as an improvement over the usual linear money demands equations suggested by [37,60,76].

The variables of the usual money demand are: $m =$ log money stock (millions), $y =$ log real net national product Y, $p =$ log of the price deflator of Y, $rs =$ log of short term interest rate, $rl =$ log of long-term interest rate, and $RS =$ short term interest rate (without logs). Let V be the velocity of circulation of money, a version of the *quantity theory of money* says that $MV(RS) = PY$ or in logs $m + v(RS) = p + y$. Rearranging terms we can write $(m - p - y) = -v(RS)$ as a *long run money demand*.

[27] applied the *2-step approach* of [20] obtaining the following results:

1st Step:

$$\left(m - p - y\right)_t = -0.31 - 7RS_t + \hat{u}_t \tag{3a}$$

where $\hat{u}_t$ are the residuals from 1878 to 2000 of the cointegrating relationship estimated by the super-consistent ordinary least squares (OLS) estimator. The inverse of the log of velocity of circulation of money, $(m - p - y) = \log\left(\frac{M}{PY}\right) = -v(RS)$, is $I(1)$ and the short run interest rate (RS) is also $I(1)$. Therefore, since the error term $\hat{u}_t$ is *stable and significant* it is $I(0)$, see conditions (e) and (f) of Theorem 1 below. Equation (3a), or (3b), is the first example of *nonlinear cointegration* given by;

$$\frac{M}{PY} = \exp(-0.31 - 7RS + \hat{u})\,. \tag{3b}$$

Similar nonlinear cointegrating relationships based on long run money demand equations are recently estimated by [3].

[22] and [27] showed that, even if OLS might not be a consistent estimator (see [95]) when the errors of (3a), (3b) are nonlinear, the OLS estimates of (3a) and the NLS estimates of (4) in 1-step are very similar.

2nd Step:

$$\begin{aligned}(1-L)\left(m-p\right)_t &= 0.45\,(1-L)\left(m-p\right)_{t-1}\\ &\quad-(1-L)^2\left(m-p\right)_{t-2}-0.60\,(1-L)\,p_t\\ &\quad+0.39\,(1-L)\,p_{t-1}-0.021\,(1-L)\,rs_t \qquad (4)\\ &\quad-0.062\left(1-L^2\right)rl_t-2.55\left(\hat{u}_{t-1}-0.2\right)\hat{u}_{t-1}^2\\ &\quad+0.005+3.7\,(D1+D3)+\hat{\varepsilon}_t\end{aligned}$$

where D_1 and D_3 are dummy variables for the two world wars. The second nonlinear characteristic of model (4), apart from the nonlinear cointegration relationship, comes from the fact that the $\hat{u}_{t-1}$ term enters in a *cubic polynomial form* as a particular nonlinear error correction (NEC) model, see also Sect. 3.3 in [17,18,95] discuss the inconsistencies derived from the 2-step approach OLS estimator in the context of nonlinear smooth transition error correction model. However, Monte Carlo simulations should be done to identify the type of nonlinearities that create series biases and inconsistencies using the 2-step estimation of NEC models.

A Nonlinear Version of Granger Representation Theorem

To justify this type of nonlinear models, we need to generalize the linear notions of temporal memory based on the linear ARIMA concepts of integration, usually $I(1)$ and $I(0)$, to nonlinear measures of dependence. Several generalizations have been proposed in the literature, as we will see later on. Our first definition is motivated from asymptotic theory, more concretely from functional central limit theorems (FCLT). See Subsect. "A Nonlinear Version of Granger Representation Theorem" for alternative definitions.

FCLT-Based Definition of $I(0)$: *A sequence $\{m_t\}$ is $I(0)$ if the "high level" condition that m_t verifies a FCLT is satisfied, i. e.*

$$T^{-1/2}\sum_{t=1}^{[Tr]} m_t \xrightarrow{d} B(r)$$

where $B(r)$ is a Brownian motion, see [73,74,99].

Definition 1 (Strong mixing) Let $\{v_t\}$ be a sequence of random variables. Let $\Im_s^t \equiv \sigma(v_s,\ldots,v_t)$ be the generated sigma-algebra. Define the α-mixing coefficients

$$\alpha_m \equiv \sup_t \sup_{\left\{F\in\Im_{-\infty}^{t},\,G\in\Im_{t+m}^{\infty}\right\}} \left|\mathrm{P}\left(G\cap F\right)-\mathrm{P}\left(G\right)\mathrm{P}\left(F\right)\right|$$

The process $\{v_t\}$ is said to be strong mixing (also α-mixing) if $\alpha_m \to 0$ as $m\to\infty$. If $\alpha_m \leqslant m^{-a}$ we say that $\{v_t\}$ is strong mixing of size $-a$.

Definition 2 (NED) Let $\{w_t\}$ be a sequence of random variables with $E\left\{w_t^2\right\} < \infty$ for all t. It is said that $\{w_t\}$ is NED on the underlying sequence $\{v_t\}$ of size $-a$ if $\phi\left(n\right)$ is of size $-a$, where $\phi\left(n\right)$ given by

$$\sup\left\|w_t - E_{t-n}^{t+n}\left(w_t\right)\right\|_2 \equiv \phi\left(n\right)$$

where $E_{t-n}^{t+n}(w_t) = E(w_t|v_{t-n},\ldots,v_{t+n})$ and $\|\cdot\|_2$ is the L_2 norm of a random vector, defined as $E^{1/2}\left|\cdot\right|^2$ where $|\cdot|$ denotes the Euclidean norm.

Weak-Dependence-Based Definition of $I(0)$: *A sequence is $I(0)$ if it is NED on an underlying α-mixing sequence $\{v_t\}$ but the sequence $\{x_t\}$ given by $x_t = \sum_{s=1}^{t} w_s$ is not NED on$\{v_t\}$. In this case, we will say that x_t is $I(1)$.*

Notice that if x_t is $I(1)$ then Δx_t is $I(0)$. This definition excludes $I(-1)$ series as $I(0)$, like $z_t = e_t - e_{t-1}$ for α-mixing sequences e_t, since in this case $\sum_{s=1}^{t} z_s$ is α-mixing.

Definition 3 Two $I(1)$ sequences $\{y_t\}$ and $\{x_t\}$ are (linearly) cointegrated with cointegrating vector $[1,-\beta]'$, if $y_t - \beta x_t$ is NED on a particular α-mixing sequence but $y_t - \delta_{12}x_t$ is not NED for $\delta_{12}\neq\beta$.

Theorem 1 (Granger's Representation Theorem, see [31]) *Consider the nonlinear correction model (NEC) for the (2×1) vector $X_t = (y_t, x_t)'$, given by*

$$\Delta X_t = \Psi 1 \Delta X_{t-1} + F(X_{t-1}) + \upsilon_t \quad . \qquad (5)$$

Assume that:

(a) *$\upsilon_t = (\upsilon_{y_t}, \upsilon_{x_t})'$ is α-mixing of size $-s/(s-2)$ for $s>2$*
(b) *$\sum_t v_t$ is not NED on α-mixing sequence*
(c) *$E\left\|v_t\right\|^2 \leqslant \Delta_v$*
(d) *$F(X_{t-1}) = J(Z_{t-1})$, where $Zt \equiv y_t - \beta x_t$ and $J(\cdot)$ is a continuously differentiable function, which satisfies a generalized Lipschitz condition of Lemma 2 of [31].*
(e) *Let $SR\left(\Psi 1\right) < 1$, where $SR(M)$ is the spectral radius of the matrix M, and*
(f) *for some fixed $\delta\in(0,1)$*

$$SR\begin{pmatrix}\Psi 1 & \nabla_z J(Z)\\ \beta'\Psi 1 & I_r + \beta'\nabla_z J(z)\end{pmatrix} \leqslant 1-\delta\,.$$

The above conditions ensure that;

(i) *ΔX_t and Z_t are simultaneously NED on the α-mixing sequence (v_t, u_t), where $u_t = v_{y,t} - \beta' v_{x,t}$; and*
(ii) *X_t is $I(1)$.*

This theorem gives sufficient conditions for cointegrated variables to be generated by a nonlinear error correction model.

Single-Equation Parametric NEC Models with Linear Cointegration

Consider the following NEC with linear cointegration

$$\Delta y_t = \psi_1 \Delta x_t + f(z_{t-1}; \gamma) + \upsilon_t$$
$$y_t = \beta x_t + z_t \,.$$

As we said in the previous section, it is not difficult to generalize this model to include other variables, lags and cointegrating relations. Consider two independent α-mixing sequences $\{a_t\}$ and $\{\upsilon_t\}$ with a zero mean. Then the following three equations represent the DGP,

$$x_t = x_{t-1} + a_t \tag{6a}$$

$$z_t = z_{t-1} + \phi_1 \Delta x_t + f(z_{t-1}, \gamma) + \nu_t \tag{6b}$$

$$y_t = \beta x_t + z_t \tag{6c}$$

where the nonlinear function $f(z_{t-1}, \gamma)$ form the nonlinear error correction term and $\beta \neq 0$. Notice that x_t is $I(1)$ by construction from (6a). Notice the similarity between Eqs. (6b) and the linear error correction of Eq. (2).

If we can ensure that z_t is NED then y_t is also $I(1)$ and linearly cointegrated with x_t, where the cointegration relationship, $y_t - \beta x_t$, is linear. If we apply the difference operator to (6c) and substitute in (6b) we obtain (7),

$$\Delta y_t = (\beta + \phi_1)\Delta x_t + f(z_{t-1}, \gamma) + \nu_t \tag{7}$$

which is a nonlinear error correction model with linear cointegration, with $\psi_1 = \beta + \phi_1$. For the sake of simplicity, and without loss of generality, we impose a *common factor restriction* so that $\phi_1 = 0$ on (7) obtaining,

$$z_t = z_{t-1} + f(z_{t-1}, \gamma) + \nu_t \tag{8}$$

and then ψ_1 equals β, the cointegration parameter, and therefore (8) is a nonlinear extension of the Dickey–Fuller equation used in unit root testing. The errors of the cointegration relation are given by $z_{t-1} = y_{t-1} - \beta x_{t-1}$, and the OLS residuals are given by $\hat{z}_{t-1} = y_{t-1} - \hat{\beta} x_{t-1}$, where $\hat{\beta}$ is the value of β estimated in the OLS regression (6c). Substituting z_t by $\hat{z}_{t-1}$ in (8) we obtain a nonlinear version of Engle and Granger's cointegration test (cf. [20]).

Differentiating (8) with respect to z_{t-1} we obtain

$$\frac{\mathrm{d}}{\mathrm{d}z_{t-1}} z_t = 1 + \frac{\mathrm{d}}{\mathrm{d}z_{t-1}} f(z_{t-1}, \gamma)$$

and therefore, our boundedness condition (see assumptions (e) and (f) of Theorem 1) is $-1 < \frac{\mathrm{d}}{\mathrm{d}z_{t-1}} z_t < 1$, or $-2 < \frac{\mathrm{d}f(z_{t-1},\gamma)}{\mathrm{d}z_{t-1}} < 0$ (models (6b), (7) and (8) are error correcting), which is sufficient to ensure that the series z_t is near epoch dependent (NED) and therefore y_t and x_t are cointegrated, see [26,27] and [31].

We discuss now few alternative nonlinear error correction (or equilibrium correction) functions $f(.)$ that could generate the series z_t from the system (6a) to (6c).

NEC Model 1: Arctan, [32]

$$f(z, \delta_1, \delta_2, \gamma_2) = -\gamma_2 \arctan(\delta_1 z + \delta_2) \quad \text{for } \gamma_2 > 0 \,.$$

NEC Model 2: Rational Polynomial, [27]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = -\gamma_2 \left((z + \delta_1)^3 + \delta_2\right) / \left((z + \delta_3)^2 + \delta_4\right) \quad \text{for } \gamma_2 > 0 \,.$$

In the first two models, the derivatives are in the desired region (satisfy assumptions (e) and (f)) for appropriate values of some of the parameters but not for all. However, within the class of rational polynomials the model considered can satisfy the condition on the absolute value of the derivative, see [27] and [30,32]. Other empirical examples of nonlinear error correction models are [11,22,29,33,49,61,72].

An important body of the literature has focused on *threshold models*, see [5,6,39,48,54,75,94], among others.

NEC Model 3: Switching Exponential, [30,32]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = \gamma_2 (\exp(-\delta_1 z) - \delta_2) I_{\{z \geqslant 0\}} + \gamma_2 (\delta_4 - \exp(\delta_3 z)) I_{\{z < 0\}} \,,$$

where $I_{\{S\}}$ is the characteristic function of the set S, $\gamma_2 > 0$, $\delta_1 > 0$ and δ_3.

NEC Model 4: Regime Switching Error Correction Models, [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^{3} 1(z \in R_s)\gamma_s z \,,$$

where $1(.)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

NEC Model 5: Random Regime Switching Error Correction Models, [6] and [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^{3} 1(z + \eta \in R_s)\gamma_s z \,,$$

where $1(.)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

Another important literature is related to *smooth transition error correction models*, see [50,90,92].

NEC Model 6: Smooth Transition Error Correction Models, [102] and [92]

$$f\,(z,\delta_1,\delta_2,\gamma_2)=\sum_{s=1}^{3}h(z)\gamma_s z$$

$$h(z)=\left\{\begin{array}{ll}1-L_1(z), & s=1\\ L_1(z)-L_2(z), & s=2\\ L_3(z), & s=3\end{array}\right\}$$

$$\text{where}\quad L_s(z)=(1+\exp\{-\gamma(z-c_s)\}$$

$$\text{for}\quad \gamma\succ 0\quad\text{and}\quad s=1,2\,.$$

Notice, that many smooth transition error correction models allow the nonlinear error correction function to affect all the parameters of the model, and not only the error correction term. However, in this paper we do not discuss them since they belong to a more general class of time varying models which is out of the context of this survey, see for example [50,101,102].

Nonparametric NEC Models with Linear Cointegration

[25,27] applied the semi-parametric *smoothing splines* estimation procedure of [21,96,104] to the estimation of the unknown nonlinear error correction function of Eq. (4). They found that in the long run money demand of the UK, see Eq. (3a), there are either multiple equilibria, or a threshold error correction with two attraction points (two equilibria); one negative and equal to -0.05 and one positive and equal to 0.2. They suggest estimating those unknown thresholds using a cubic polynomial parametric functional form, see Eq. (4). Notice that the corresponding cubic polynomial error correction term, $-2.55\,(\hat{u}_{t-1}-0.2)\,\hat{u}^2_{t-1}$, identifies perfectly one of the thresholds, the one that is equal to 0.2. The second threshold could be obtained from the roots of the polynomial. Other empirical examples of threshold error correction models are [5,10,48,54]. In fact [10] used a similar nonparametric approach to estimate the nonlinear error correction function, but instead of using smoothing splines they used the Nadaraya–Watson kernel estimator discussed in [55].

Nonlinear Cointegration

In the recent years several proposals have been considered to extend linear cointegration and linear error correction of Granger [42] to a nonlinear framework. One possibility is to allow for a NEC model in the Granger's representation. We have discussed such an approach in the previous section. Alternatively, one may consider a nonlinear cointegration relation.

Despite the fact that many macroeconomic and financial time series dynamics are nonlinear, there are still today relatively few useful analytical tools capable of assessing the dependence and persistence behavior of nonlinear time series appropriately (cf. [50]). This problem is even more accentuated by the fact that traditional measures of dependence, such as autocorrelations and periodograms, may be inappropriate when the underlying time series is nonlinear and/or non-Gaussian. Then, it is generally accepted that new measures of dependence have to be defined in order to develop a new concept of nonlinear cointegration. We have already discussed measures based on FCLT and on NED concepts. We shall explore several alternative measures in this section. All the measures considered can be grouped in measures of conditional mean dependence or in distributional dependence. Higher order conditional moments, other than the mean, can of course be considered. In any case, we shall use the general terminology *extended memory* and *short memory* to indicate a nonlinear persistence and non-persistence process, respectively (cf. [44]).

Once a concept of nonlinear persistence is introduced, a general definition of nonlinear cointegration is as follows. We say that two"extended memory" series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = f(y_t,x_t)$ is short memory. This definition is more appropriate when dealing with distributional persistence, and it is perhaps too general to be fully operative. Identification problems arise in this general context, as noted by many authors, so one should restrict the class of functions f to avoid such identification problems. [46] considered functions of the form $z_t=g(y_t)-h(x_t)$, and estimate g and h nonparametrically by means of the Alternating Conditional Expectations (ACE) algorithm. See also [40] for a related approach. It is still an open problem the theoretical justification of these nonparametric estimation procedures.

A less ambitious approach is to consider transformations of the form $z_t=y_t-f(x_t)$. This framework is especially convenient with conditional mean persistence measures. We review the existing measures in the next section.

Nonlinear Measures of Memory

As already discussed by [46], a generalization of linear cointegration to a nonlinear set-up goes through proper extensions of the linear concepts of $I(0)$ and $I(1)$. We introduce in this section alternative definitions of nonlinear $I(0)$ and $I(1)$ processes. We first focus on conditional mean persistence, we shall discuss distributional dependence at the end of this section. Define the conditional mean function $E(y_{t+h}|I_t)$, where $I_t = (x_t, x_{t-1}, \dots)$ is the conditioning set at time t. [44] defines the Short Memory in Mean (SMM) and Extended Memory in Mean (EMM) concepts as follows.

Definition 4 (SMM and EMM) $\{y_t\}$ is said to be SMM if for all t, $M(t,h) = E(y_{t+h}|I_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E\,|M(t,h)-\mu|^2 < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity, for all t. If $\{y_t\}$ does not satisfy the previous condition is called EMM.

Note that to be mathematically precise in Definition 4, we should specify that $\{y_t\}$ is SMM or EMM with respect to $\{x_t\}$. Referring to this definition, [44] considered the case $x_t = y_t$. [46] replaced the name of EMM by long memory in mean, and [40] denoted EMM and SMM by nonlinear integrated (NLI) and nonlinear integrated of order zero (NLI(0)), respectively. As noted by [44] the concepts of SMM and EMM are related to a kind of "mixing in mean" property, more precisely to the concept of mixingale, see [16].

[23] introduced the pairwise equivalent measures of the previous concepts, which, although weaker, are more operative because they only involve finite-dimensional random variables.

Definition 5 (PSMM and PEMM) $\{y_t\}$ is said to be Pairwise SMM (PSMM) if for all t, $m(t,h) = E(y_{t+h}|x_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E|m(t,h)-\mu|^2 < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity, for all t. If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM (PEMM).

From the previous definitions and the law of iterated expectations, we easily observe that a process SMM is PSMM. The reciprocal is false. There exist processes which are PSMM but not SMM, although they are rare in practice.

We now discuss generalizations of the usual autocovariances and crosscovariances to a nonlinear framework. These generalizations were introduced by [23]. It is well-known that in the presence of nonlinearity (or non-Gaussianity) the autocovariances do not characterize the dependence and the practitioner needs more reliable measures such as the pairwise regression functions $m(t,h)$. In general, inference on these functions involves nonparametric estimation with bandwidth choices, hampering their application to practical situations. By a measure-theoretic argument, the regression function $m(t,h)$ can be characterized by the integrated regression function $\gamma_{t,h}(x)$ given by

$$\begin{aligned}\gamma_{t,h}(x) &= E\left[(y_{t+h}-\mu_t)1(x_t \leqslant x)\right]\\ &= E\left[m(t,h)1(x_t \leqslant x)\right],\end{aligned}$$

where the second equality follows by the law of iterated expectations. The measures $\gamma_{t,h}(x)$ are called the Integrated Pairwise Regression Functions (IPRF), see [24]. Extensions to other weight functions different from the indicator weight $1(x_t \leqslant x)$ are possible. The integrated measures of dependence $\gamma_{t,h}(x)$ are useful for testing interesting hypotheses in a nonlinear time series framework and, unlike $m(t,h)$, they do not need of smoothing estimation and are easily estimated by the sample analogue. Moreover, they characterize the pairwise versions of the concepts introduced by [44], making these concepts more operative. First we need a definition, a norm $\|\|$ is nondecreasing if for all f and g with $|f(x)| \leqslant |g(x)|$ for all x, it holds that $\|f\| \leqslant \|g\|$. Associated to the norm $\|\|$, we define the distance $d(f,g) = \|f-g\|$. Usual nondecreasing norms are the L_2 norm and the supremum norm.

Definition 6 (PSMM$_d$ and PEMM$_d$) $\{y_t\}$ is said to be Pairwise SMM relative to d (PSMM$_d$) if for all t, $\left\|\gamma_{t,h}(x)\right\|$, $h > 0$, tends to zero as h becomes large for any t. More precisely, $\left\|\gamma_{t,h}(x)\right\| < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity for all t. If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM relative to d (PEMM$_d$).

Theorem 2 (Relationship between PSMM and PSMM$_d$ [23]) *If the norm $\|\|$ is non-decreasing and $E\left\{y_t^2\right\} < \infty$ for all t, then $\{y_t\}$ is PSMM if and only if $\{y_t\}$ is PSMM$_d$.*

Based on these concepts we define nonlinear cointegration as follows. We say that two PEMM$_d$ series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = y_t - f(x_t)$ is PSMM$_d$.

In analogy with the linear world and based on the results in [1,78], a possible nonparametric characterization of nonlinear cointegration can be based on the rates of convergence of $\gamma_{t,h}(x)$ and $\gamma^x_{t,h}(x) = E\left[(x_{t+h}-\mu_t)1(x_t \leqslant x)\right]$ as h diverges to infinity. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in long-run non-linearly forecasting y_t.

Similarly, we can define distributional measures of persistence and nonlinear cointegration. [45] defines a persistent process in distribution using the bivariate and marginal densities at different lags. [41] considered parametric and nonparametric methods for studying serial distributional dependence. In the nonparametric case [45] consider series expansions estimators for the nonlinear canonical analysis of the series. These authors apply their results to study the dynamics of the inter-trade durations of the Alcatel-stock on the Paris Borse and find evidence of nonlinear strong persistence in distribution.

Regarding distributional dependence, we formalize a definition given in [45]. Let $f_{t,h}(y,x)$, $k_{t+h}(y)$ and $g_t(x)$ be, respectively, the bivariate and marginal densities of y_{t+h} and x_t. To define persistence in distribution one can use the Hellinger distance

$$H_{t,h} = \iint \left| f_{t,h}^{1/2}(y,x) - k_{t+h}^{1/2}(y) g_t^{1/2}(x) \right|^2 \mathrm{d}y\mathrm{d}x \,,$$

and define Pairwise Short Memory in Distribution (PSMD) according to the decay of $H_{t,h}$ to zero as h diverges to infinity. Alternative definitions can be given in terms of other divergence measures or distances, see [51] and references therein. This approach is explored in [1], who define nonlinear cointegration using mutual information measures.

Persistence in distribution is related to mixing concepts. In fact, uniformly in t, $H_{t,h} \leqslant 2\alpha(h)$, where $\alpha(h)$ is certain α-mixing coefficient, see [23] for details.

The aforementioned measures of nonlinear distributional dependence need of smoothing estimation, e. g. kernel estimators. Similarly to the case of conditional mean measures, we can avoid smoothing by means of the integrated measures of dependence

$$\begin{aligned} \eta_{t,h}(y,x) &= \operatorname{cov}(1(y_{t+h} \leqslant y), 1(x_t \leqslant x)) \\ &= F_{t,h}(y,x) - K_{t+h}(y) G_t(x) \,, \end{aligned}$$

where $F_{t,h}(y,x)$, $K_{t+h}(y)$ and $G_t(x)$ are, respectively, the bivariate and marginal cumulative distribution functions (cdf) of y_{t+h} and x_t. The measures $\eta_{t,h}(y,x)$ can bt estimated at different lags by using the sample analogue, i. e. the empirical distribution functions. Similar definitions to Definition 6 can be given for distributional persistence based on $\eta_{t,h}(y,x)$. See [23] for further generalizations and definitions. Definitions of nonlinear cointegration can be formulated along the lines in [1]. For instance, we can say that two persistent (in distribution) series y_t and x_t are nonlinear cointegrated (in distribution) if

$$\lim_{\tau\to\infty} \left\| \frac{\eta_{t,\tau}(\cdot)}{\eta_{t,\tau}^{x}(\cdot)} - \beta \right\| = 0$$

for all τ, t, $\tau \leqslant t$, where $\eta_{t,h}^{x}(y,x)$ is defined as $\eta_{t,h}(y,x)$ but replacing y_t by x_t there, β is a real number and $\|\|$ a suitable norm.

Integration and Cointegration Based on Order Statistics

[9,36,47] have considered rank based unit roots test to avoid the extreme sensitivity of usual test like the Dickey–Fuller type test to presence of nonlinearities and outliers, see [19] for an overview of the problems of unit root tests in general contexts. [2] suggested a range unit root test (RUR) based on the first differences of the ranges of a series. The range is the difference between the maximum and the minimum taken by a series at any given point in time. Therefore, the difference of the ranges is a measure of records. Counting the number of new records is an interesting way of distinguishing between stationary and nonstationary series since the frequency of new records vanishes faster for stationary series than for series containing unit roots. They have shown that this RUR test is robust to monotonic transformations, distributions of the errors and many structural breaks and additive outliers.

[8] suggests using the differences between the sequences of ranks. If there is no cointegration the sequence of ranks tends to diverge, while under cointegration they evolve similarly. [34] consider a record counting cointegration test (RCC) based on the synchronicity of the jumps between cointegrated series. They suggest a test statistic based on counting the number of jumps (records) that simultaneously occur in both series. Certainly, those series that are cointegrated have a much larger number of simultaneous jumps in the ranges of the series. They show that the cointegration test based on RCC is robust to monotonic nonlinearities and many structural breaks and does not require a prior estimation of the nonlinear or linear cointegrating function. There is a large literature on the effects of structural breaks and outliers on unit root and on cointegration testing but it is out of the scope of this paper, see for example the references in [62].

Another cointegration test robust to nonlinearities and structural breaks is based on induced order statistics. In particular [35] consider that two series y_t and x_t are cointegrated (either linear or nonlinear) if the corresponding induced order series are plotted around the 45° line. Their test-statistic compares the two induced order series by comparing their corresponding empirical distributions, the empirical distribution of y_t and the empirical distribution of y_t induced by x_t, using the Kolmogorov–Smirnov type statistic.

Parametric Nonlinear Cointegration

As previously discussed, an important body of research has focused on nonlinear cointegration relations. The model used being a nonlinear cointegration regression or a nonlinear regression model with integrated regressors. References on this line of research include [12,13,82,84]. [93] study smooth transition regressions with integrated regressors. An example considered by these authors is the nonlinear extension of (1b)

$$y_t = \alpha x_t + \delta x_t g(x_t - c, \gamma) + v_t ,$$

where $g(x - c, \gamma) = -1/(1 + e^{-\gamma(x-c)})$ is the logistic function. Under the restriction $\delta = 0$ the latter model reduces to the linear model in (1b), see [14] for linearity tests under this framework. Other examples of parametric nonlinear cointegration are given by Eqs. (3a), (3b) and by the exponential model of [68], see also the references given in [19]. Those models are a particular case of the more general *nonlinear parametric cointegration* model of the form,

$$y_t = f(x_t, \beta) + v_t ,$$

where x_t is $px1$ vector of $I(1)$ regressors, v_t is zero-mean stationary error term, and $f(x_t, \beta)$ an smooth function of the process x_t, known up to the finite-dimensional parameter vector β.

At least, there are two different asymptotic justifications within these nonlinear cointegration models. In the classical asymptotic theory (e. g. [82,84]) all the existing literature has been confined to the case $p = 1$, although some extensions to single-index type regressions have been studied, see [83]. The main reason for the restriction to the univariate case is that commonly used asymptotic techniques are not appropriate for the case $p > 1$, e. g. asymptotics based on local times are not available for $p > 2$. Intrinsic to this problem is the non-recurrent property of the p-variate Brownian motion when $p > 2$. On the other hand, the triangular array asymptotics used in [93] allow for a general $p > 0$.

In the classical asymptotic theory, the properties of estimators of β, e. g. the nonlinear least squares estimator (NLSE), depend on the specific class of functions where $f(x_t, \beta)$ belongs. Commonly used classes are integrable functions, asymptotic homogeneous functions or exponential functions, see [82]. The rate of convergence of the NLSE is class-specific and, in some cases, involves random scaling. In the triangular array asymptotic theory of [93] the distribution theory of estimators of β, e. g. rates of convergence, does not depend on the specific class of functions.

Several authors exploit the previous asymptotic results on β to develop tests of nonlinear cointegration in this parametric framework. In [15] the so-called KPSS test is applied to the parametric residuals. Since the resulting limiting distribution depends on nuisance parameters, these authors implement the test with the assistance of a subsampling procedure as a smoothing device.

Nonparametric Nonlinear Cointegration

Nonparametric estimates of nonlinear cointegration relations were already computed by [46], but it has not been until the recent works by [69,70,88] that a nonparametric estimation theory for nonstationary processes has been developed. [88] considered a theory based on local time arguments, whereas [69,70] used the theory of null recurrent Markov processes. A comparison of both methodologies is discussed in [4], where near-integrated nonparametric asymptotics are studied.

More specifically, these authors estimate the transfer function $f(x_t)$ in the nonlinear regression model

$$y_t = f(x_t) + v_t ,$$

where the series y_t and x_t are univariate observed nonstationary processes and v_t is a non-observed stationary process. [70] study the nonparametric kernel estimation of $f(x_t)$ as

$$\hat{f}(x) = \frac{\sum_{t=0}^{n} y_t K_{x,h}(x_t)}{\sum_{t=0}^{n} K_{x,h}(x_t)} ,$$

where $K_{x,h}(x_t) = h^{-1} K((y - x)/h)$, K is a kernel function and h is a bandwidth parameter. These authors investigate the asymptotic theory for $\hat{f}(x)$ under some regularity conditions and different assumptions on the dependence relation between x_t and v_t. Especially convenient for the nonlinear cointegration framework are those assumptions that allow for dependence between x_t and v_t. The family of nonstationary processes considered by these authors is the class of the so-called β-null recurrent Markov processes satisfying a restriction on the tail distribution of the recurrence time. The class is large enough to contain the random walk, unit-root processes as well as other nonlinear nonstationary processes. It is shown that the nonparametric estimation theory is different to that in the stationary case, with slower rates of convergences, as expected. This new nonparametric asymptotic theory opens the door for future developments in inferences in nonlinear cointegration models.

Future Directions

This chapter has provided a selected overview of the available nonlinear extensions of this concept. While in the linear set-up there exists a complete theory and set of tools for studying the cointegration problem, it has been made clear that a nonlinear version of this theory possesses nontrivial challenges. A first natural nonlinear extension is to allow for a NEC model but still a linear cointegration regression. On the other hand, one can consider a nonlinear regression cointegration equation. It has been recognized that an extension of the concept of linear cointegration to a nonlinear set-up needs of appropriate extensions of the linear concepts of $I(0)$ and $I(1)$ to nonlinear time series (cf. [46]). Several extensions have been provided and discussed. We recommend operative integrated measures of dependence, since they are simple to estimate and avoid smoothing of the data, which can be a challenging problem when dealing with nonstationary variables (cf. [69]). An important line of future research is the development of inferential procedures for nonlinear cointegration based on these new integrated measures. This is currently investigated by the authors.

There is a large evidence of empirical applications in economics and finance where nonlinearities are found in nonstationary contexts. However, given the difficulty of the theory involved, only few papers provide a sound justification of the empirical use of cointegration regressions (nonlinear cointegration) in nonlinear frameworks. The difficulties analyzing nonlinear time series models within a stationary and ergodic frameworks are substantially enhanced in nonstationary contexts. In particular, the classical asymptotic theory for nonlinear transformations of nonstationary variables becomes case-dependent (i. e. depends on the specific class of functions), and the available results are confined to the univariate case. A challenging and important line of research deals with the extension of this theory to multivariate frameworks.

Recently, an important step towards the development of a nonlinear cointegration theory has been accomplished by the nonparametric estimation theory of [69,70]. The application of this theory to inference in nonlinear cointegrated models is not fully explored yet. Residual-based tests for testing nonlinear cointegration, such as the so-called KPSS test (cf. [73]), can be constructed using nonparametric residuals. Moreover, model specification tests for nonlinear parametric cointegration can be based on the comparison between parametric and nonparametric fits. Finally, in a recent unpublished lecture, Clive Granger suggested to extend the concept of cointegration to quantiles. These are promising lines of future research which deserve serious attention in the economics literature.

Bibliography

Primary Literature

1. Aparicio F, Escribano A (1999) Information-theoretic analysis of serial correlation and cointegration. Stud Nonlinear Dyn Econ 3:119–140
2. Aparicio F, Escribano A, Sipols AE (2006) Range unit root (RUR) tests: Robust against nonlinearities, error distributions, structural breaks and outliers. J Time Ser Anal 27:545–576
3. Bae Y, de Jong RM (2005) Money demand function estimation by nonlinear cointegration. Working Paper, Department of Economics, Ohio State University
4. Bandi FM (2004) On persistente and nonparametric estimation (with an application to stock return predictability). Unpublished manuscript
5. Balke NS, Fomby TB (1997) Threshold cointegration. Int Econ Rev 38:627–645
6. Bec F, Rahbek A (2004) Vector equilibrium correction models with nonlinear discontinuous adjustments. Econ J 7:628–651
7. Boswijk HP (1995) Efficient inference on cointegration parameters in structural error correction models. J Econ 69:113–158
8. Breitung J (2001) Rank tests for nonlinear cointegration. J Bus Econ Stat 19:331–340
9. Breitung J, Goriéroux C (1997) Rank tests for unit roots. J Econ 81:7–27
10. Breitung J, Wulff C (2001) Nonlinear error correction and the efficient market hypothesis: The case of german dual-class shares. Ger Econ Rev 2:419–434
11. Burgess SM (1992) Nonlinear dynamics in a structural model of employment. J Appl Econ 7:101–118
12. Chang Y, Park JY (2003) Index models with integrated time series. J Econ 114:73–16
13. Chang Y, Park JY, Phillips PCB (2001) Nonlinear econometric models with cointegrated and deterministically trending regressors. Econ J 4:1–36
14. Choi I, Saikkonen P (2004) Testing linearity in cointegrating smooth transition regressions. Econ J 7:341–365
15. Choi I, Saikkonen P (2005) Tests for nonlinear cointegration. Unpublished manuscript
16. Davidson J (1994) Stochastic limit theory. Oxford U. P., New York
17. de Jong RM (2001) Nonlinear estimation using estimated cointegrating relations. J Econ 101:109–122
18. de Jong RM (2002) Nonlinear minimization estimators in the presence of cointegrating relations. J Econ 110:241–259
19. Dufrénot G, Mignon V (2002) Recent developments in nonlinear cointegration with applications to macroeconomics and finance. Kluwer, Boston
20. Engle RF, Granger CWJ (1987) Co-integration and error correction: Representation, estimation, and testing. Econetrica 55:251–276
21. Engle RF, Granger CWJ, Rice J, Weiss A (1986) Semiparametric estimates of the relationship between weather and electricity sales. J Am Stat Assoc 81:310–320
22. Ericsson NR, Hendry DF, Prestwich KM (1998) The demand for broad money in the United Kingdom, 1878–1993. Scand J Econ 100:289–324
23. Escanciano JC, Hualde J (2005) Persistence and long memory in nonlinear time series. Unpublished manuscript
24. Escanciano JC, Velasco C (2006) Testing the martingale difference hypothesis using integrated regression functions. Comput Stat Data Anal 51:2278–2294
25. Escribano A (1986) Non-Linear error-correction: The case of money demand in the UK (1878–1970), ch IV. Ph.D. Dissertation, University of California, San Diego
26. Escribano A (1987) Error-Correction systems: Nonlinear adjustment to linear long-run relationships. Core Discussion Paper 8730, C.O.R.E

27. Escribano A (2004) Nonlinear error correction: The case of money demand in the UK (1878–2000). Macroecon Dyn 8:76–116
28. Escribano A, Aparicio F (1999) Cointegration: Linearity, nonlinearity, outliers and structural breaks. In: Dahiya SB (ed) The current state of economic science. Spellbound Publications, pp 383–408
29. Escribano A, Granger CWJ (1998) Investigating the relationship between gold and silver prices. J Forecast 17:81–107
30. Escribano A, Mira S (1997) Nonlinear error correction models. Working Paper 97-26. Universidad Carlos III de Madrid
31. Escribano A, Mira S (2002) Nonlinear error correction models. J Time Ser Anal 23:509–522
32. Escribano A, Mira S (2007) Specification of nonlinear error correction models: a simulation study. Mimeo, Universidad Carlos III de Madrid
33. Escribano A, Pfann GA (1998) Non-linear error correction, asymmetric adjustment and cointegration. Econ Model 15:197–216
34. Escribano A, Sipols AE, Aparicio F (2006) Nonlinear cointegration and nonlinear error correction: Applications of tests based on first differences of ranges. Commun Stat Simul Comput 35:939–956
35. Escribano A, Santos MT, Sipols AE (2008) Testing for cointegration using induced order statistics. Comput Stat 23:131–151
36. Fotopoulos SB J, Ahn SK (2001) Rank based Dickey-Fuller tests statistics. J Time Ser Anal 24:647–662
37. Friedman M, Schwartz A (1982) Monetary trends in the united sates and the United Kingdom: Their relation to income, prices, and interest rates, 1867–1975. University of Chicago Press, Chicago
38. Gonzalo J (1994) Five alternative methods of estimating long run equilibrium relationships. J Econ 60:1–31
39. Gonzalo J, Pitarakis J-Y (2006) Threshold cointegrating relationships. Oxford Bull Econ Stat 68:813–833
40. Gouriéroux C, Jasiak J (1999) Nonlinear persistence and copersistence. Unpublished manuscript
41. Gouriéroux C, Jasiak J (2002) Nonlinear autocorrelograms: An application to inter-trade durations. J Time Ser Anal 23:127–154
42. Granger CWJ (1981) Some properties of time series data and their use in econometric model specification. J Econ 16:121–130
43. Granger CWJ (1986) Developments in the study of cointegrated variables. Oxford Bull Econ Stat 48:213–228
44. Granger CWJ (1995) Modelling nonlinear relationships between extended-memory variables. Econetrica 63:265–279
45. Granger CWJ (2002) Long memory, volatility, risk and distribution. Unpublished manuscript
46. Granger CWJ, Hallman J (1991) Long memory series with attractors. Oxford Bull Econ Stat 53:11–26
47. Granger CWJ, Hallman J (1991) Nonlinear transformations of integrated time series. J Time Ser Anal 12:207–224
48. Granger CWJ, Lee TH (1989) Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. J Appl Econ 4:145–159
49. Granger CWJ, Swanson N (1996) Further developments in the study of cointegrated variables. Oxford Bull Econ Stat 58:537–553
50. Granger CWJ, Teräsvirta T (1993) Modelling nonlinear economic relationships. Oxford University Press, New York
51. Granger CWJ, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. J Time Ser Anal 25:649–669
52. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
53. Hansen B, Phillips PCB (1990) Estimation and inference in models of cointegration: A simulation study. Adv Econ 8:225–248
54. Hansen B, Seo B (2002) Testing for two-regime threshold cointegration in vector error correction models. J Econ 110:293–318
55. Härdle W (1990) Applied nonparametric regression. Econetric Society Monographs, vol 19. Cambridge University Press, Cambridge
56. Hargreaves CP (1994) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
57. Hargreaves CP (1994) A review of methods of estimating cointegrating relationships. In: Hargreaves CP (ed) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
58. Haug AA (2002) Testing linear restrictions on cointegrating vectors: Sizes and powers of wald and likelihood ratio tests in finite samples. Econ Theory 18:505–524
59. Hendry DF (1995) Dynamic econometrics. Oxford University Press, Oxford
60. Hendry DF, Ericsson N (1983) Assertion without empirical basis: An econometric appraisal of 'monetary trends in the … the United Kingdom' by Friedman M, Schwartz AJ. In: Monetary trends in the United Kingdom. Bank of England Panel of Academic Consultants, Paper 22, 45–101
61. Hendry DF, Ericsson NR (1991) An econometric analysis of the uk money demand in 'monetary trends in the united states and the United Kingdom' by Friedman M, Schwartz AJ. Am Econ Rev 81:8–38
62. Hendry DF, Massmann M (2007) Co-breaking: Recent advances and a synopsis of the literature. J Bus Econ Stat 25:33–51
63. Inder B (1993) Estimating long-run relationships in economics: A comparison of different approaches. J Econ 57:53–68
64. Johansen S (1988) Statistical analysis of cointegration vectors. J Econ Dyn Control 12:231–54
65. Johansen S (1991) Estimation and hypothesis testing of cointegration vectors, in gaussian vector autoregressive models. Econetrica 59:1551–1580
66. Johansen S (1992) Cointegration in partial systems and the efficiency of single-equation analysis. J Econ 52:389–402
67. Juselius K (2006) The cointegrated VAR model: Methodology and applications. Oxford University Press, Oxford
68. Jusmah A, Kunst RM (2008) Modelling macroeconomic subaggregates: An application of non-linear cointegration. Economics Series, Institute for Advanced Studies. Macroecon Dyn 12:151–171
69. Karlsen HA, Tjøstheim D (2001) Nonparametric estimation in null recurrent time series. Ann Stat 29:372–416
70. Karlsen HA, Myklebust T, Tjøstheim D (2007) Nonparametric estimation in a nonlinear cointegration type model. Ann Stat 35(1):252–299
71. Kitamura Y, Phillips PCB (1995) Efficient IV estimation in nonstationary regression: An overview and simulation study. Econ Theory 11:1095–1130

72. Kunst RM (1992) Threshold cointegration in interest rates. Working Paper, Institute for Advanced Studies, Vienna
73. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. J Econ 54:159–178
74. Lo AW (1991) Long-term memory in stock market prices. Econometrica 59:1279–1313
75. Lo M, Zivot E (2001) Threshold cointegration and nonlinear adjustment to the law of one price. Macroecon Dyn 5:533–576
76. Longbottom A, Holly S (1985) Econometric methodology and monetarism: Professor Friedman and Professor Hendry on the demand for money, Discussion Paper No. 131. London Business School
77. Lütkephol H (2005) New introduction to multiple time series analysis. Springer
78. Marmol F, Escribano A, Aparicio FM (2002) Instrumental variable interpretation of cointegration with inference results for fractional cointegration. Econ Theory 18:646–672
79. Park JY (1992) Canonical cointegrating regressions. Econometrica 60:119–143
80. Park JY, Phillips PCB (1988) Statistical inference in regressions with integrated process: Part 1. Econ Theory 4:468–498
81. Park JY, Phillips PCB (1989) Statistical inference in regressions with integrated process: Part 2. Econ Theory 4:95–132
82. Park JY, Phillips PCB (1989) Asymptotics for nonlinear transformations of integrated time series. Econ Theory 15:269–298
83. Park JY, Phillips PCB (2000) Nonstationary binary choice. Econometrica 68:1249–1280
84. Park JY, Phillips PCB (2001) Nonlinear regressions with integrated time series. Econometrica 69:117–161
85. Phillips PCB (1991) Optimal inference in cointegrated systems. Econometrica 59:283–306
86. Phillips PCB, Hansen BE (1990) Statistical inference in instrumental variable regression with *I*(1) processes. Rev Econ Stud 57:99–125
87. Phillips PCB, Loretan M (1991) Estimating long-run economic equilibria. Rev Econ Stud 59:407–436
88. Phillips PCB, Park JY (1998) Nonstationary density estimation and kernel autoregression. Unpublished manuscript
89. Robinson PM, Gerolimetto M (2006) Instrumental variables estimation of stationary and nonstationary cointegrating regressions. Econ J 9:291–306
90. Rothman P, van Dijk D, Franses PH (2001) A multivariate star analysis of the relationship between money and output. Macroecon Dyn 5:506–532
91. Saikkonen P (1991) Asymptotically efficient estimation of cointegration regressions. Econ Theory 7:1–21
92. Saikkonen P (2005) Stability results for nonlinear error correction models. J Econ 127:69–81
93. Saikkonen P, Choi I (2004) Cointegrating smooth transition regressions. Econ Theory 20:301–340
94. Seo MH (2006) Bootstrap testing for the null of no cointegration in a threshold vector error correction model. J Econ 134:129–150
95. Seo MH (2007) Estimation of nonlinear error correction models. Unpublished manuscript
96. Silverman BW (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. J Roy Stat Soc Ser B 47:1–52
97. Sims CA (1972) Money, income, and causality. Am Econ Rev 62:540–552
98. Stock JH (1987) Asymptotic properties of least squares estimation of cointegrating vectors. Econometrica 55:1035–1056
99. Stock JH (1994) Deciding between *I*(1) and *I*(0). J Econ 63:105–131
100. Stock JH, Watson MW (1993) A simple estimator of cointegration vectors in higher order integrated systems. Econometrica 61:783–820
101. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) Handbook of applied economic statistics. Marcel Dekker, New York, pp 507–552
102. Teräsvirta T, Eliasson AC (2001) Non-linear error correction and the uk demand for broad money, 1878–1993. J Appl Econ 16:277–288
103. Velasco C (2006) Semiparametric estimation of long-memory models. In: Patterson K, Mills TC (eds) Palgrave handbook of econometrics, vol 1. Econometric Theory. MacMillan, Palgrave, pp 353–395
104. Wahba G (1975) Smoothing noisy data with spline functions. Num Math 24:309-63–75
105. Watson M (1994) Large sample estimation and hypothesis testing, vector autoregression and cointegration. In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol IV. Elsevier Science, Amsterdam

Books and Reviews

Franses PH, Terasvirta T (eds)(2001) Special issue on nonlinear modeling of multivariate macroeconomic relations. Macroecon Dyn 5(4)

Banerjee A, Dolado JJ, Galbraith JW, Hendry DF (1993) Co-integration, error correction and the econometric analysis of non-stationary data. Oxford University Press, Oxford

Bierens HJ (1981) Robust methods and asymptotic theory in nonlinear econometrics. Lecture Notes in Economics and Mathematical Systems, vol 192. Springer, Berlin

Clements MP, Hendry DF (1999) Forecasting non-stationary economic time series. MIT Press, Cambridge

Dhrymes P (1998) Time series, unit roots, and cointegration. Academic Press, San Diego

Enders W (1995) Applied econometric time series. Wiley

Engle RF, Granger CWJ (eds) (1991) Long-run economic relationships: Readings in cointegration. Oxford University Press, Oxford

Franses PH, van Dijk D (2000) Nonlinear time series models in empirical finance. Cambridge University Press, Cambridge

Gonzalo J, Dolado JJ, Marmol F (2001) A primer in cointegration. In: Baltagui BH (ed) A companion to theoretical econometrics. Blackwell, New York, ch 30

Granger CWJ (2001) Overview of nonlinear macroeconometric empirical models. Macroecon Dyn 5:466–481

Granger CWJ (2007) Some thoughts on the past and future of cointegration. Mimeo

Hatanaka M (1996) Time series-based econometrics. Oxford University Press, Oxford

Maddala GS, Kim I-M (1998) Unit roots, cointegration and structural change. Cambridge University Press, Cambridge

Phillips PCB (1986) Understanding spurious regressions in econometrics. J Econ 33:311–340

Econometrics: Panel Data Methods

JEFFREY M. WOOLDRIDGE
Department of Economics, Michigan State University,
East Lansing, USA

Article Outline

Glossary

Panel data Data on a set of cross-sectional units followed over time.

Unobserved effects Unobserved variables that affect the outcome which are constant over time.

Fixed effects estimation An estimation method that removes the unobserved effects, implying that the unobserved effects can be arbitrarily related to the observed covariates.

Correlated random effects An approach to modeling where the dependence between the unobserved effects and the history of the covariates is parametrically modeled. The traditional random effects approach is a special case under the assumption that he unobserved effects are independent of the covariates.

Average partial effect The partial effect of a covariate averaged across the distribution of the unobserved effects.

Definition of the Subject

Panel data consist of repeated observations over time on the same set of cross-sectional units. These units can be individuals, firms, schools, cities, or any collection of units one can follow over time. Special econometric methods have been developed to recognize and exploit the rich information available in panel data sets. Because the time dimension is a key feature of panel data sets, issues of serial correlation and dynamic effects need to be considered. Further, unlike the analysis of cross-sectional data, panel data sets allow the presence of systematic, unobserved differences across units that can be correlated with observed factors whose effects are to be measured. Distinguishing between persistence due to unobserved heterogeneity and that due to dynamics in the underlying process is a leading challenge for interpreting estimates from panel data models.

Panel data methods are the econometric tools used to estimate parameters compute partial effects of interest in nonlinear models, quantify dynamic linkages, and perform valid inference when data are available on repeated cross sections. For linear models, the basis for many panel data methods is ordinary least squares applied to suitably transformed data. The challenge is to develop estimators assumptions with good properties under reasonable assumptions, and to ensure that statistical inference is valid. Maximum likelihood estimation plays a key role in the estimation of nonlinear panel data models.

Introduction

Many questions in economics, especially those with foundations in the behavior of relatively small units, can be empirically studied with the help of panel data. Even when detailed cross-sectional surveys are available, collecting enough information on units to account for systematic differences is often unrealistic. For example, in evaluating the effects of a job training program on labor market outcomes, unobserved factors might affect both participation in the program and outcomes such as labor earnings. Unless participation in the job training program is randomly assigned, or assigned on the basis of observed covariates, cross-sectional regression analysis is usually unconvincing. Nevertheless, one can control for this individual heterogeneity – including unobserved, time-constant human capital – by collecting a panel data set that includes data points both before and after the training program.

Some of the earliest econometric applications of panel data methods were to the estimation of agricultural production functions, where the worry was that unobserved inputs – such as soil quality, technical efficiency, or managerial skill of the farmer – would generally be correlated with observed inputs such as capital, labor, and amount of land. Classic examples are [31,45].

The nature of unobserved heterogeneity was discussed early in the development of panel data models. An important contribution is [46], which argued persuasively that in applications with many cross-sectional units and few time periods, it always makes sense to treat unit-specific heterogeneity as outcomes of random variables, rather than parameters to estimate. As Mundlak made clear, for economic applications the key issue is whether the unobserved heterogeneity can be assumed to be independent, or at least uncorrelated, with the observed covariates. [25]

developed a testing framework that can be used, and often is, to test whether unobserved heterogeneity is correlated with observed covariates. Mundlak's perspective has had a lasting impact on panel data methods, and his insights have been applied to a variety of dynamic panel data models with unobserved heterogeneity.

The 1980s witnessed an explosion in both methodological developments and applications of panel data methods. Following the approach in [15,16,46], and [17] provided a unified approach to linear and nonlinear panel data models, and explicitly dealt with issues of inference in cases where full distributions were not specified. Dynamic linear models, and the problems they pose for estimation and inference, were considered in [4]. Dynamic discrete response models were analyzed in [29,30]. The hope in estimating dynamic models that explicitly contain unobserved heterogeneity is that researchers can measure the importance of two causes for persistence in observed outcomes: unobserved, time-constant heterogeneity and so-called *state dependence*, which describes the idea that, conditional on observed and unobserved factors, the probability of being in a state in the current time period is affected by last period's state.

In the late 1980s and early 1990s, researchers began using panel data methods to test economic theories such as rational expectations models of consumption. Unlike macro-level data, data at the individual or family level allows one to control for different preferences, and perhaps different discount rates, in testing the implications of rational expectations. To avoid making distributional assumptions on unobserved shocks and heterogeneity, researchers often based estimation on conditions on expected values that are implied by rational expectations, as in [40].

Other developments in the 1990s include studying standard estimators under fewer assumptions – such as the analysis in [53] of the fixed effects Poisson estimator under distributional misspecification and unrestricted serial dependence – and the development of estimators in nonlinear models that are consistent for parameters under no distributional assumptions – such as the new estimator proposed in [33] for the panel data censored regression model.

The past 15 years has seen continued development of both linear and nonlinear models, with and without dynamics. For example, on the linear model front, methods have been proposed for estimating models where the effects of time-invariant heterogeneity can change over time – as in [2]. Semiparametric methods for estimating production functions, as in [48], and dynamic models, as in the dynamic censored regression model in [34], have been developed. Flexible parametric models, estimated by maximum likelihood, have also been proposed (see [57]).

Many researchers are paying closer attention to estimation of partial effects, and not just parameters, in nonlinear models – with or without dynamics. Results in [3] show how partial effects, with the unobserved heterogeneity appropriately averaged out, can be identified under weak assumptions.

The next several sections outline a modern approach to panel data methods. Section "Future Directions" provides an account of more recent advances, and discusses where those advances might head in the future.

Overview of Linear Panel Data Models

In panel data applications, linear models are still the most widely used. When drawing data from a large population, random sampling is often a realistic assumption; therefore, we can treat the observations as independent and identically distributed outcomes. For a random draw i from the population, the linear panel data model with additive heterogeneity can be written as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}\,, \quad t = 1, \ldots, T\,, \tag{1}$$

where T is the number of time periods available for each unit and t indexes time periods. The time periods are often years, but the span between periods can be longer or shorter than a year. The distance between any two time periods need not be the same, although different spans can make it tricky to estimate certain dynamic models. As written, Eq. (1) assumes that we have the same time periods available for each cross-sectional unit. In other words, the panel data set is *balanced*.

As in any regression analysis, the left-hand-side variable is the dependent variable or the response variable. The terms η_t, which depend only only time, are treated here as parameters. In most microeconometric applications, the cross-sectional sample size, denoted N, is large – often very large – compared with T. Therefore, the η_t can be estimated precisely in most cases. Almost all applications should allow for aggregate time effects as captured by η_t. Including such time effects allows for secular changes in the economic environment that affect all units in the same way (such as inflation or aggregate productivity). For example, in studying the effects of school inputs on performance using school-level panel data for a particular state, including η_t allows for trends in statewide spending along with separate, unrelated trends in statewide test performance. It could be that, say, real spending rose at the same time that the statewide standardized tests were made easier; a failure to account for such aggregate trends could lead to a spurious association between performance and spending. Only occasionally are the η_t the focus of a panel data analysis, but it is sometimes interesting to study the pattern of aggregate changes once the covariates contained in the $1 \times K$ vector $\mathbf{x}_{it}$ are netted out.

The parameters of primary interest are contained in the $K \times 1$ vector $\boldsymbol{\beta}$, which contains the coefficients on the set of explanatory variables. With the presence of η_t in (1), $\mathbf{x}_{it}$ cannot include variables that change only over time. For example, if y_{it} is a measure of labor earnings for individual i in year t for a particular state in the US, $\mathbf{x}_{it}$ cannot contain the state-level unemployment rate. Unless interest centers on how individual earnings depend on the state-level unemployment rate, it is better to allow for different time intercepts in an unrestricted fashion: this way, any aggregate variables that affect each individual in the same way are accounted for without even collecting data on them. If the η_t are restricted to be functions of time – for example, a linear time trend – then aggregate variables can be included, but this is always more restrictive than allowing the η_t to be unrestricted.

The composite error term in (1), $c_i + u_{it}$, is an important feature of panel data models. With panel data, it makes sense to view the unobservables that affect y_{it} as consisting of two parts: the first is the time-constant variable, c_i, which is often called an *unobserved effect* or *unit-specific heterogeneity*. This term aggregates all factors that are important for unit i's response that do not change over time. In panel data applications to individuals, c_i is often interpreted as containing cognitive ability, family background, and other factors that are essentially determined prior to the time periods under consideration. Or, if i indexes different schools across a state, and (1) is an equation to see if school inputs affect student performance, c_i includes historical factors that can affect student performance and also might be correlated with observed school inputs (such as class sizes, teacher competence, and so on). The word "heterogeneity" is often combined with a qualifier that indicates the unit of observation. For example, c_i might be "individual-specific heterogeneity" or"school-specific heterogeneity". Often in the literature c_i is called a "random effect" or "fixed effect", but these labels are not ideal. Traditionally, c_i was considered a random effect if it was treated as a random variable, and it was considered a fixed effect if it was treated as a parameter to estimate (for each i). The flaws with this way of thinking are revealed in [46]: the important issue is not whether c_i is random, but whether it is correlated with $\mathbf{x}_{it}$.

The sequence of errors $\{u_{it}\colon t = 1, \ldots, T\}$ are specific to unit i, but they are allowed to change over time. Thus, these are the time-varying unobserved factors that affect y_{it}, and they are often called the *idiosyncratic errors*. Because u_{it} is in the error term at time t, it is important to know whether these unobserved, time-varying factors are uncorrelated with the covariates. It is also important to recognize that these idiosyncratic errors can be serially correlated, and often are.

Before treating the various assumptions more formally in the next subsection, it is important to recognize the asymmetry in the treatment of the time-specific effects, η_t, and the unit-specific effects, c_i. Language such as "both time and school fixed effects are included in the equation" is common in empirical work. While the language itself is harmless, with large N and small T it is best to view the time effects, η_t, as parameters to estimate because they can be estimated precisely. As already mentioned earlier, viewing c_i as random draws is the most general, and natural, perspective.

Assumptions and Estimators for the Basic Model

The assumptions discussed in this subsection are best suited to cases where random sampling from a (large) population is realistic. In this setting, it is most natural to describe large-sample statistical properties as the cross-sectional sample size, N, grows, with the number of time periods, T, fixed.

In describing assumptions in the model (1), it probably makes more sense to drop the i subscript in (1) to emphasize that the equation holds for an entire population. Nevertheless, (1) is useful for emphasizing which factors change i, or t, or both. It is sometimes convenient to subsume the time dummies in $\mathbf{x}_{it}$, so that the separate intercepts η_t need not be displayed.

The traditional starting point for studying (1) is to rule out correlation between the idiosyncratic errors, u_{it}, and the covariates, $\mathbf{x}_{it}$. A useful assumption is that the sequence of explanatory variables $\{\mathbf{x}_{it}\colon t = 1, \ldots, T\}$ is *contemporaneously exogenous conditional on* c_i:

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0\,, \quad t = 1, \ldots, T\,. \tag{2}$$

This assumption essentially defines $\boldsymbol{\beta}$ in the sense that, under (1) and (2),

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i\,, \tag{3}$$

so the β_j are partial effects holding fixed the unobserved heterogeneity (and covariates other than x_{tj}). Strictly speaking, c_i need not be included in the conditioning set in (2), but including it leads to the useful Eq. (3). Plus, for purposes of stating assumptions for inference, it is convenient to express the contemporaneous exogeneity assumption as in (2).

Unfortunately, with a small number of time periods, $\boldsymbol{\beta}$ is not identified by (2), or by the weaker assumption $\mathrm{Cov}(\mathbf{x}_{it}, u_{it}) = \mathbf{0}$. Of course, if c_i is assumed to be uncorrelated with the covariates, that is $\mathrm{Cov}(\mathbf{x}_{it}, c_i) = \mathbf{0}$ for any t, then the composite error, $v_{it} = c_i + u_{it}$ is uncorrelated with $\mathbf{x}_{it}$, and then $\boldsymbol{\beta}$ is identified and can be consistently estimated by a cross section regression using a single time period t, or by using pooled regression across t. (See Chaps. 7 and 10 in [55] for further discussion.) But one of the main purposes in using panel data is to allow the unobserved effect to be correlated with time-varying $\mathbf{x}_{it}$.

Arbitrary correlation between c_i and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ is allowed if the sequence of explanatory variables is *strictly exogenous conditional on c_i*,

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0\,, \quad t = 1, \ldots, T\,, \tag{4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i\,. \tag{5}$$

Clearly, assumption (4) implies (2). Because the entire history of the covariates is in (4) for all t, (4) implies that $\mathbf{x}_{ir}$ and u_{it} are uncorrelated for all r and t, including $r = t$. By contrast, (2) allows arbitrary correlation between $\mathbf{x}_{ir}$ and u_{it} for any $r \neq t$. The strict exogeneity assumption (4) can place serious restrictions on the nature of the model and dynamic economic behavior. For example, (4) can never be true if $\mathbf{x}_{it}$ contains lags of the dependent variable. Of course, (4) would be false under standard econometric problems, such as omitted time-varying variables, just as would (2). But there are important cases where (2) can hold but (4) might not. If, say, a change in u_{it} causes reactions in future values of the explanatory variables, then (4) is generally false. In applications to the social sciences, the potential for these kind of "feedback effects" is important. For example, in using panel data to estimate a firm-level production function, a shock to production today (captured by changes in u_{it}) might affect the amount of capital and labor inputs in the next time period. In other words, u_{it} and $\mathbf{x}_{i,t+1}$ would be correlated, violating (4).

How does assumption (4) (or (5)) identify the parameters? In fact, it only allows estimation of coefficients on time-varying elements of $\mathbf{x}_{it}$. Intuitively, because (4) puts no restrictions on the dependence between c_i and $\mathbf{x}_i$, it is not possible to distinguish between the effect of a time-constant observable covariate and that of the unobserved effect, c_i. For example, in an equation to describe the amount of pension savings invested in the stock market, c_i might include innate of tolerance for risk, assumed to be fixed over time. Once c_i is allowed to be correlated with any observable covariate – including, say, gender – the effects of gender on stock market investing cannot be identified because gender, like c_i, is constant over time. Mechanically, common estimation methods eliminate c_i along with any time-constant explanatory variables. (What is meant by "time-varying" x_{itj} is that for at least some i, x_{itj} changes over time. For some units i, x_{itj} might be constant). When a full set of year intercepts – or even just a linear time trend – is included, the effects of variables that increase by the same amount in each period – such as a person's age – cannot be included in $\mathbf{x}_{it}$. The reason is that the beginning age of each person is indistinguishable from c_i, and then, once the initial age is know, each subsequent age is a deterministic – in fact, linear – function of time.

Perhaps the most common method of estimating $\boldsymbol{\beta}$ (and the η_t) is so-called *fixed effects (FE)* or *within* estimation. The FE estimator is obtained as a pooled OLS regression on variables that have had the unit-specific means removed. More precisely, let $\ddot{y}_{it} = y_{it} - T^{-1}\sum_{r=1}^{T} y_{ir} = y_{it} - \bar{y}_i$ be the deviation of y_{it} from the average over time for unit i, $\bar{y}_i$ and similarly for $\ddot{\mathbf{x}}_{it}$ (which is a vector). Then,

$$\ddot{y}_{it} = \ddot{\eta}_t + \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}\,, \quad t = 1, \ldots, T\,, \tag{6}$$

where the year intercepts and idiosyncratic errors are, of course, also demeaned. Consistency of pooled OLS (for fixed T and $N \to \infty$) applied to (6) essentially requires rests on $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = \sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}' u_{it}) = \mathbf{0}$, which means the error u_{it} should be uncorrelated with $\mathbf{x}_{ir}$ for all r and t. This assumption is implied by (4). A rank condition on the demeaned explanatory variables is also needed. If $\ddot{\eta}_t$ is absorbed into $\ddot{\mathbf{x}}_{it}$, the condition is *rank* $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}) = K$, which rules out time constant variables and other variables that increase by the same value for all units in each time period (such as age).

A different estimation method is based on an equation in first differences. For $t > 1$, define $\Delta y_{it} = y_{it} - y_{i,t-1}$, and similarly for the other quantities. The first-differenced equation is

$$\Delta y_{it} = \delta_t + \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}\,, \quad t = 2, \ldots, T\,, \tag{7}$$

where $\delta_t = \eta_t - \eta_{t-1}$ is the change in the intercepts. The *first-difference* (FD) estimator is pooled OLS applied to (7). Any element x_{ith} of $\mathbf{x}_{it}$ such that Δx_{ith} is constant for all i and t (most often zero) drops out, just as in FE estimation. Assuming suitable time variation in the covariates, $E(\Delta\mathbf{x}_{it}'\Delta u_{it}) = \mathbf{0}$ is sufficient for consistency. Naturally, this assumption is also implied by assumption (4).

Whether FE or FD estimation is used – and it is often prudent to try both approaches – inference about $\boldsymbol{\beta}$

can and generally should be be made fully robust to heteroksedasticity and serial dependence. The robust asymptotic variance of both FE and FD estimators has the so-called"sandwich" form, which allows the vector of idiosyncratic errors, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$, to contain arbitrary serial correlation and heteroskedasticity, where the conditional covariances and variances can depend on $\mathbf{x}_i$ in an unknown way. For notational simplicity, absorb dummy variables for the different time periods into $\mathbf{x}_{it}$. Let $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ denote the fixed effects estimator and $\widehat{\ddot{\mathbf{u}}}_i = \ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i \hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ the $T \times 1$ vector of fixed effects residuals for unit i. Here, $\ddot{\mathbf{X}}_i$ is the $T \times K$ matrix with t^{th} row $\ddot{\mathbf{x}}_{it}$. Then a fully robust estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ is

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) = \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1} \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \widehat{\ddot{\mathbf{u}}}_i \widehat{\ddot{\mathbf{u}}}_i' \ddot{\mathbf{X}}_i \right) \cdot \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1} , \quad (8)$$

where it is easily seen that $\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i = \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}_{it}$ and the middle part of the sandwich consists of terms $\widehat{\ddot{u}}_{ir} \widehat{\ddot{u}}_{it} \ddot{\mathbf{x}}_{ir}' \ddot{\mathbf{x}}_{it}$ for all $r, t = 1, \ldots, T$. See Chap. 10 in [55] for further discussion. A similar expression holds for $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$ but where the demeaned quantities are replaced by first differences.

When $T = 2$, it can be shown that the FE and FD estimation and inference about $\boldsymbol{\beta}$ are identical. If $T > 2$, the procedures generally differ. If (4) holds and $T > 2$, how does one choose between the FE and FD approaches? Because both are consistent and $\sqrt{N}$-asymptotically normal, the only way to choose is from efficiency considerations. Efficiency of the FE and FD estimators hinges on second moment assumptions concerning the idiosyncratic errors. Briefly, if $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$, then the FE estimator is efficient. Practically, the most important implication of this assumption is that the idiosyncratic errors are serially uncorrelated. But they should also be homoskedastic, which means the variances can neither depend on the covariates nor change over time. The FD estimator is efficient if the errors in (7) are serially uncorrelated and homoskedasticity, which can be stated as $E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i' | \mathbf{x}_i) = E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i') = \sigma_e^2 \mathbf{I}_{T-1}$, where $e_{it} = u_{it} - u_{i,t-1}$ and $\Delta \mathbf{u}_i$ is the $T - 1$ vector of first-differenced errors. These two sets of conditions – that $\{u_{it} \colon t = 1, \ldots, T\}$ is a serially uncorrelated sequence (for FE to be efficient) versus $\{u_{it} \colon t = 1, \ldots, T\}$ is a random walk (for FD to be efficient) – represent extreme cases. Of course, there is much in between. In fact, probably neither condition should be assumed to be true, which is a good argument for robust inference. More efficient estimation can be based on generalized method of moments (GMM – see Chap. 8 in [55] – or minimum distance estimation, as in [16]).

It is good practice to compute both FE and FD estimates to see if they differ in substantive ways. It is also helpful to have a formal test of the strict exogeneity assumption that is easily computable and that maintains only strict exogeneity under the null – in particular, that takes no stand on whether the FE or FD estimator is asymptotically efficient. Because lags of covariates can always be included in a model, the primary violation of (4) that is of interest is due to feedback. Therefore, it makes sense to test that $\mathbf{x}_{i,t+1}$ is uncorrelated with u_{it}. Actually, let $\mathbf{w}_{it}$ be a subset of $\mathbf{x}_{it}$ that is suspected of failing the strict exogeneity assumption, and consider the augmented model

$$y_{it} = \eta_t + \mathbf{x}_{it} \boldsymbol{\beta} + \mathbf{w}_{i,t+1} \boldsymbol{\delta} + c_i + u_{it} , \quad t = 1, \ldots, T-1 . \quad (9)$$

Under the null hypothesis that $\{\mathbf{x}_{it} \colon t = 1, \ldots, T\}$ is strictly exogenous, $H_0 \colon \boldsymbol{\delta} = \mathbf{0}$, and this is easily tested using fixed effects (using all but the last time period) or first differencing (where, again, the last time period is lost). It makes sense, as always, to make the test fully robust to serial correlation and heteroskedasticity. This test may probably has little power for detecting contemporaneous endogeneity, that is, correlation between $\mathbf{w}_{it}$ and u_{it}.

A third common approach to estimation of unobserved effects models is so-called *random effects* estimation. RE estimation differs from FE and FD by leaving c_i in the error term and then accounting for its presence via generalized least squares (GLS). Therefore, the exogeneity requirements of the covariates must be strengthened. The most convenient way of stating the key random effects (RE) assumption is

$$E(c_i | \mathbf{x}_i) = E(c_i) , \quad (10)$$

which ensures that every element of $\mathbf{x}_i$ – that is, all explanatory variables in all time periods – is uncorrelated with c_i. Together with (4), (10) implies

$$E(v_{it} | \mathbf{x}_i) = 0 , \quad t = 1, \ldots, T , \quad (11)$$

where $v_{it} = c_i + u_{it}$ is the composite error. Condition (11) is the key condition for general least squares methods that exploit serial correlation in v_{it} to be consistent (although zero correlation would be sufficient). The random effects estimator uses a special structure for the variance-covariate matrix of $\mathbf{v}_i$, the $T \times 1$ vector of composite errors. If $E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$ and c_i is uncorrelated with each u_{it} (as implied by assumption (4)), then

$$\mathrm{Var}(\mathbf{v}_i) = \begin{pmatrix} \sigma_c^2+\sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2+\sigma_u^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_c^2 \\ \sigma_c^2 & \cdots & \sigma_c^2 & \sigma_c^2+\sigma_u^2 \end{pmatrix} . \quad (12)$$

Both σ_c^2 and σ_u^2 can be estimated after, say, preliminary estimation by pooled OLS (which is consistent under (11)) – see, for example, Chap. 10 in [55] – and then a feasible GLS is possible. If (12) holds, along with the system homoskedasticity assumption $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i) = \mathrm{Var}(\mathbf{v}_i)$, then feasible GLS is efficient, and the inference is standard. Even if $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i)$ is not constant, or $\mathrm{Var}(\mathbf{v}_i)$ does not have the random effects structure in (12), the RE estimator is consistent provided (11) holds (Again, this is with N growing and T fixed). Therefore, although it is still not common, a good case can be made for using robust inference – that is, inference that allows an unknown form of $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i)$ – in the context of random effects. The idea is that the RE estimator can be more efficient than pooled OLS even if (12) fails, yet inference should not rest on (12). Chapter 10 in [55] contains the sandwich form of the estimator.

Under the key RE assumption (11), $\mathbf{x}_{it}$ can contain time-constant variables. In fact, one way to ensure that the omitted factors are uncorrelated with the key covariates is to include a rich set of time-constant controls in $\mathbf{x}_{it}$. RE estimation is most convincing when many good time-constant controls are available. In some applications of RE, the key variable of interest does not change over time, which is why FE and FD cannot be used. (Methods proposed in [26] can be used when some covariates are correlated with c_i, but enough others are assumed to be uncorrelated with c_i).

Rather than eliminate c_i using the FE or FD transformation, or assuming (10) and using GLS, a different approach is to explicitly model the correlation between c_i and $\mathbf{x}_i$. A general approach is to write

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\lambda} + a_i , \quad (13)$$

$$E(a_i) = 0 \quad \text{and} \quad E(\mathbf{x}_i' a_i) = \mathbf{0} , \quad (14)$$

where $\boldsymbol{\lambda}$ is a $TK \times 1$, vector of parameters. Equations (13) and (14) are definitional, and simply define the population linear regression of c_i on the entire set of covariates, $\mathbf{x}_i$. This representation is due to [16], and is an example of a *correlated random effects* (CRE) model. The uncorrelated random effects model occurs when $\boldsymbol{\lambda} = \mathbf{0}$.

A special case of (13) was used in [46], assuming that each $\mathbf{x}_{ir}$ has the same set of coefficients. Plus, [46] actually used conditional expectations (which is unnecessary but somewhat easier to work with):

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i \quad (15)$$

$$E(a_i|\mathbf{x}_i) = 0 , \quad (16)$$

where recall that $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$. This formulation conserves on degrees of freedom, and extensions are useful for nonlinear models.

Plugging (15) into the original equation gives

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it} , \quad (17)$$

where ψ is absorbed into the time intercepts. The composite error $a_i + u_{it}$ satisfies $E(a_i + u_{it}|\mathbf{x}_i) = 0$, and so pooled OLS or random effects applied to (17) produces consistent, $\sqrt{N}$-asymptotically normal estimators of all parameters, including $\boldsymbol{\xi}$. In fact, if the original model satisfies the second moments ideal for random effects, then so does (17). Interesting, both pooled OLS and RE applied to (17) produce the fixed effects estimate of $\boldsymbol{\beta}$ (and the η_t). Therefore, the FE estimator can be derived from a correlated random effects model. (Somewhat surprisingly, the same algebraic equivalence holds using Chamberlain's more flexible device. Of course, the pooled OLS estimator is not generally efficient, and [16] shows how to obtain the efficient minimum distance estimator. See also Chap. 11 in [55]).

One advantage of Eq. (17) is that it provides another interpretation of the FE estimate: it is obtained by holding fixed the time averages when obtaining the partial effects of each x_{itj}. This results in a more convincing analysis than not controlling for systematic differences in the levels of the covariates across i.

Equation (17) has other advantages over just using the time-demeaned data in pooled OLS: time-constant variables can be included in (17), and the resulting equation gives a simple, robust way of testing whether the time-varying covariates are uncorrelated with It is helpful to write the original equation as

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it} , \quad t = 1, \ldots, T , \quad (18)$$

where $\mathbf{g}_t$ is typically a vector of time period dummies but could instead include other variables that change only over time, including linear or quadratic trends, $\mathbf{z}_i$ is a vector of time-constant variables, and $\mathbf{w}_{it}$ contains elements that vary across i and t. It is clear that, in comparing FE to RE estimation, $\boldsymbol{\gamma}$ can play no role because it cannot be estimated by FE. What is less clear, but also true, is that the coefficients on the aggregate time variables, $\boldsymbol{\eta}$, cannot be included in any comparison, either. Only the $M \times 1$ estimates of $\boldsymbol{\delta}$, say $\hat{\boldsymbol{\delta}}_{\mathrm{FE}}$ and $\hat{\boldsymbol{\delta}}_{\mathrm{RE}}$, can be compared. If $\hat{\boldsymbol{\eta}}_{\mathrm{FE}}$ and

$\hat{\boldsymbol{\eta}}_{\text{RE}}$ are included, the asymptotic variance matrix of the difference in estimators has a nonsingularity in the asymptotic variance matrix. (In fact, RE and FE estimation only with aggregate time variables are identical.) The Mundlak equation is now

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + u_{it}\,, \quad t = 1, \ldots, T\,, \tag{19}$$

where the intercept is absorbed into $\mathbf{g}_t$. A test of the key RE assumption is $H_0\colon \boldsymbol{\xi} = \mathbf{0}$ is obtained by estimating (19) by RE, and this equation makes it clear there M restrictions to test. This test was described in [5,46] proposed the robust version. The original test based directly on comparing the RE and FE estimators, as proposed in [25], it more difficult to compute and not robust because it maintains that the RE estimator is efficient under the null.

The model in (19) gives estimates of the coefficients on the time-constant variables $\mathbf{z}_i$. Generally, these can be given a causal interpretation only if

$$E(c_i|\mathbf{w}_i, \mathbf{z}_i) = E(c_i|\mathbf{w}_i) = \psi + \bar{\mathbf{w}}_i\boldsymbol{\xi}\,, \tag{20}$$

where the first equality is the important one. In other words, $\mathbf{z}_i$ is uncorrelated with c_i once the time-varying covariates are controlled for. This assumption is too strong in many applications, but one still might want to include time-constant covariates.

Before leaving this subsection, it is worth point out that generalized least squares methods with an unrestricted variance-covariance matrix can be applied to every estimating equation just presented. For example, after eliminating c_i by removing the time averages, the resulting vector of errors, $\ddot{\mathbf{u}}_i$, can have an unrestricted variance matrix. (Of course, there is no guarantee that this matrix is the same as the variance matrix conditional on the matrix of time-demeaned regressors, $\ddot{\mathbf{X}}_i$.) The only glitch in practice is that $\text{Var}(\ddot{\mathbf{u}}_i)$ has rank $T-1$, not T. As it turns out, GLS with an unrestricted variance matrix for the original error vector, $\mathbf{u}_i$, can be implemented on the time-demeaned equation with any of the T time periods dropped. The so-called *fixed effects GLS* estimates are invariant to whichever equation is dropped. See [41] or [37] for further discussion. The initial estimator used to estimate the variance covariance matrix would probably be the usual FE estimator (applied to all time periods).

Feasible GLS can be applied directly the first differenced equation, too. It can also be applied to (19), allowing the composite errors $a_i + u_{it}$, $t = 1, \ldots, T$, to have an unrestricted variance-covariance matrix. In all cases, the assumption that the conditional variance matrix equals the unconditional variance can fail, and so one should use fully robust inference even after using FGLS. Chapter 10 in [55] provides further discussion. Such options are widely available in software, sometimes under the rubric of *generalized estimating equations* (GEE). See, for example, [43].

Models with Heterogeneous Slopes

The basic model described in the previous subsection introduces a single source of heterogeneity in the additive effect, c_i. The form of the model implies that the partial effects of the covariates depend on a fixed set of population values (and possibly other unobserved covariates if interactions are included in $\mathbf{x}_{it}$). It seems natural to extend the model to allow interactions between the observed covariates and time-constant, unobserved heterogeneity:

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{21}$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0\,, \quad t = 1, \ldots, T\,, \tag{22}$$

where $\mathbf{b}_i$ is $K \times 1$. With small T, one cannot precisely estimate $\mathbf{b}_i$. Instead, attention usually focuses on the *average partial effect* (APE) or *population averaged effect* (PAE). In (21), the vector of APEs is $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$, the $K \times 1$ vector of means. In this formulation, aggregate time effects are in $\mathbf{x}_{it}$. This model is sometimes called a *correlated random slopes* model – which means the slopes are allowed to be correlated with the covariates.

Generally, allowing $(c_i, \mathbf{b}_i)$ and $\mathbf{x}_i$ to be arbitrarily correlated requires $T > K+1$ – see [56]. With a small number of time periods and even a modest number of regressors, this condition often fails in practice. Chapter 11 in [55] discusses how to allow only a subset of coefficients to be unit specific. Of interest here is the question: if the usual FE estimator is applied – that is, ignoring the unit-specific slopes $\mathbf{b}_i$ – does this ever consistently estimate the APEs in $\boldsymbol{\beta}$? In addition to the usual rank condition and the strict exogeneity assumption (22), [56] shows that a simple sufficient condition is

$$\text{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \text{E}(\mathbf{b}_i) = \boldsymbol{\beta}\,, \quad t = 1, \ldots, T\,. \tag{23}$$

Importantly, condition (23) allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. It rules out correlation between idiosyncratic movements in $\mathbf{x}_{it}$ and $\mathbf{b}_i$. For example, suppose the covariates can be decomposed as $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}, t = 1, \ldots, T$. Then (23) holds if $\text{E}(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = \text{E}(\mathbf{b}_i)$. In other words, $\mathbf{b}_i$ is allowed to be arbitrarily correlated with the permanent component, $\mathbf{f}_i$. Condition (23) is similar in spirit to the key assumption in [46] for the intercept c_i: the correlation between the slopes b_{ij} and the entire history $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ is through the time averages, and not through deviations from the time averages. If $\mathbf{b}_i$ changes across i, ignoring it by using the usual FE estimator effectively puts $\ddot{\mathbf{x}}_{it}(\mathbf{b}_i - \boldsymbol{\beta})$ in the error term, which induces heteroskedasticity and serial correlation in the composite error even if the $\{u_{it}\}$ are homoskedastic and serially independent. The possible presence of this term provides another argument for making inference with FE fully robust to arbitrary conditional and unconditional second moments.

The (partial) robustness of FE to the presence of correlated random slopes extends to a more general class of estimators that includes the usual fixed effects estimator. Write an extension of the basic model as

$$y_{it} = \mathbf{g}_t \mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} , \quad t = 1, \ldots, T , \tag{24}$$

where $\mathbf{g}_t$ is a set of deterministic functions of time. A leading case is $\mathbf{g}_t = (1, t)$, so that each unit has its own time trend along with a level effect. (The resulting model is sometimes called a *random trend model*). Now, assume that the random coefficients, $\mathbf{a}_i$, are swept away be regressing y_{it} and $\mathbf{x}_{it}$ each on $\mathbf{g}_t$ for each i. The residuals, $\ddot{y}_{it}$ and $\ddot{\mathbf{x}}_{it}$, have had unit-specific trends removed, but the $\mathbf{b}_i$ are treated as constant in the estimation. The key condition for consistently estimating $\boldsymbol{\beta}$ can still be written as in (23), but now $\ddot{\mathbf{x}}_{it}$ has had more features removed at unit-specific level. When $\mathbf{g}_t = (1, t)$, each covariate has been demeaned within each unit. Therefore, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$. Of course, individually detrending the $\mathbf{x}_{it}$ requires at least three time periods, and it decreases the variation in $\ddot{\mathbf{x}}_{it}$ compared with the usual FE estimator. Not surprisingly, increasing the dimension of $\mathbf{g}_t$ (subject to the restriction $\dim(\mathbf{g}_t) < T$), generally leads to less precision of the estimator. See [56] for further discussion.

Sequentially Exogenous Regressors and Dynamic Models

The summary of models and estimators from Sect. "Overview of Linear Panel Data Models"used the strict exogeneity assumption $E(u_{it}|\mathbf{x}_i, c_i) = 0$ for all t, and added an additional assumption for models with correlated random slopes. As discussed in Sect. "Overview of Linear Panel Data Models", strict exogeneity is not an especially natural assumption. The contemporaneous exogeneity assumption $E(u_{it}|\mathbf{x}_{it}, c_i) = 0$ is attractive, but the parameters are not identified. In this section, a middle ground between these assumptions, which has been called a *sequential exogeneity assumption*, is used. But first, it is helpful to understand properties of the FE and FD estimators when strict exogeneity fails.

Behavior of Estimators Without Strict Exogeneity

Both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the strict exogeneity assumption stated in Eq. (4). But it is also pretty well known that, at least under certain assumptions, the FE estimator can be expected to have less "bias" for larger T. Under the contemporaneous exogeneity assumption (2) and the assumption that the data series $\{(\mathbf{x}_{it}, u_{it}) \colon t = 1, \ldots, T\}$ is"weakly dependent" – in time series parlance, "integrated of order zero", or $I(0)$ – then it can be shown that

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{\text{FE}} = \boldsymbol{\beta} + O(T^{-1}) \tag{25}$$

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{\text{FD}} = \boldsymbol{\beta} + O(1) ; \tag{26}$$

see Chap. 11 in [55]. In some very special cases, such as the simple AR(1) model discussed below, the "bias" terms can be calculated, but not generally.

Interestingly, the same results can be shown if $\{\mathbf{x}_{it} \colon t = 1, \ldots, T\}$ has unit roots as long as $\{u_{it}\}$ is $I(0)$ and contemporaneous exogeneity holds. However, there is a catch: if $\{u_{it}\}$ is $I(1)$ – so that the time series version of the "model" would be a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not "cointegrated"), then (25) is no longer true. On the other hand, first differencing means any unit roots are eliminated and so there is little possibility of a spurious regression. The bottom line is that using "large T" approximations such as those in (25) and (26) to choose between FE over FD obligates one to take the time series properties of the panel data seriously; one must recognize the possibility that the FE estimation is essentially a spurious regression.

Consistent Estimation Under Sequential Exogeneity

Because both the FE and FD estimators are inconsistent for fixed T, it makes sense to search for estimators that are consistent for fixed T. A natural specification for dynamic panel data models, and one that allows consistent estimation under certain assumptions, is

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i , \tag{27}$$

which says that $\mathbf{x}_{it}$ contains enough lags so that further lags of variables are not needed. When the model is written in error form, (27) is the same as

$$E(u_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = 0 , \quad t = 1, 2, \ldots, T . \tag{28}$$

Under (28), the covariates $\{\mathbf{x}_{it} \colon t = 1, \ldots, T\}$ are said to be *sequentially exogenous conditional on* c_i. Some estimation methods are motivated by a weaker version of (28), namely,

$$E(\mathbf{x}'_{is}u_{it}) = \mathbf{0}\,, \quad s = 1,\dots,t\,, \quad t = 1,\dots,T\,, \qquad (29)$$

but (28) is natural in most applications.

Assumption (28) is appealing in that it allows for finite distributed lag models as well as models with lagged dependent variables. For example, the finite distributed lag model

$$y_{it} = \eta_t + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + \cdots + \mathbf{z}_{i,t-L}\boldsymbol{\delta}_L + c_i + u_{it} \quad (30)$$

allows the elements of $\mathbf{z}_{it}$ to have effects up to L time periods after a change. With $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \dots, \mathbf{z}_{i,t-L})$, Assumption (28) implies

$$\begin{aligned} &E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, \dots, c_i) \\ &= E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, c_i) \\ &= \eta_t + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + \cdots + \mathbf{z}_{i,t-L}\boldsymbol{\delta}_L + c_i\,, \end{aligned} \qquad (31)$$

which means that the distributed lag dynamics are captured by L lags. The important difference with the strict exogeneity assumption is that sequential exogeneity allows feedback from u_{it} to $\mathbf{z}_{ir}$ for $r > t$.

How can (28) be used for estimation? The FD transformation is natural because of the sequential nature of the restrictions. In particular, write the FD equation as

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}\,, \quad t = 2,\dots,T\,. \qquad (32)$$

Then, under (29),

$$E(\mathbf{x}'_{is}\Delta u_{it}) = \mathbf{0}\,, \quad s = 1,\dots,t-1\,; \quad t = 2,\dots,T\,, \qquad (33)$$

which means any $\mathbf{x}_{is}$ with $s < t$ can be used as an instrument for the time t FD equation. An efficient estimator that uses (33) is obtained by stacking the FD equations as

$$\Delta\mathbf{y}_i = \Delta\mathbf{X}_i\boldsymbol{\beta} + \Delta\mathbf{u}_i\,, \qquad (34)$$

where $\Delta\mathbf{y}_i = (\Delta y_{i2}, \Delta y_{i3}, \dots, \Delta y_{iT})'$ is the $(T-1)\times 1$ vector of first differences and $\Delta\mathbf{X}_i$ is the $(T-1)\times K$ matrix of differences on the regressors. (Time period dummies are absorbed into $\mathbf{x}_{it}$ for notational simplicity.) To apply a system estimation method to (34), define

$$\mathbf{x}^o_{it} \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{it})\,, \qquad (35)$$

which means the valid instruments at time t are in $\mathbf{x}^o_{i,t-1}$ (minus redundancies, of course). The matrix of instruments to apply to (34) is

$$\mathbf{W}_i = \text{diag}(\mathbf{x}^o_{i1}, \mathbf{x}^o_{i2}, \dots, \mathbf{x}^o_{i,T-1})\,, \qquad (36)$$

which has $T-1$ rows and a large number of columns. Because of sequential exogeneity, the number of valid instruments increases with t.

Given $\mathbf{W}_i$, it is routine to apply generalized method of moments estimation, as summarized in [27,55]. A simpler strategy is available that can be used for comparison or as the first-stage estimator in computing the optimal weighting matrix. First, estimate a reduced form for $\Delta\mathbf{x}_{it}$ separately for each t. In other words, at time t, run the regression $\Delta\mathbf{x}_{it}$ on $\mathbf{x}^o_{i,t-1}$, $i = 1,\dots,N$, and obtain the fitted values, $\widehat{\Delta\mathbf{x}}_{it}$. Of course, the fitted values are all $1 \times K$ vectors for each t, even though the number of available instruments grows with t. Then, estimate the FD Eq. (32) by pooled IV using $\widehat{\Delta\mathbf{x}}_{it}$ as instruments (not regressors). It is simple to obtain robust standard errors and test statistics from such a procedure because the first stage estimation to obtain the instruments can be ignored (asymptotically, of course).

One potential problem with estimating the FD equation using IVs that are simply lags of $\mathbf{x}_{it}$ is that changes in variables over time are often difficult to predict. In other words, $\Delta\mathbf{x}_{it}$ might have little correlation with $\mathbf{x}^o_{i,t-1}$. This is an example of the so-called "weak instruments" problem, which can cause the statistical properties of the IV estimators to be poor and the usual asymptotic inference misleading. Identification is lost entirely if $\mathbf{x}_{it} = \boldsymbol{\lambda}_t + \mathbf{x}_{i,t-1} + \mathbf{q}_{it}$, where $\mathrm{E}(\mathbf{q}_{it}|\mathbf{x}_{i,t-1}, \dots, \mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of $\mathbf{x}_{it}$ are random walks with drift. Then, then $\mathrm{E}(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1}, \dots, \mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails. Of course, if some elements of $\mathbf{x}_{it}$ are strictly exogenous, then their changes act as their own instruments. Nevertheless, typically at least one element of $\mathbf{x}_{it}$ is suspected of failing strict exogeneity, otherwise standard FE or FD would be used.

In situations where simple estimators that impose few assumptions are too imprecise to be useful, sometimes one is willing to improve estimation of $\boldsymbol{\beta}$ by adding more assumptions. How can this be done in the panel data case under sequential exogeneity? There are two common approaches. First, the sequential exogeneity condition can be strengthened to the assumption that the conditional mean model is *dynamically complete*, which can be written in terms of the errors as

$$E(u_{it}|\mathbf{x}_{it}, y_{i,t-1}\mathbf{x}_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}, c_i) = 0\,, \quad t = 1,\dots,T\,. \qquad (37)$$

Clearly, (37) implies (28). Dynamic completeness is neither stronger nor weaker than strict exogeneity, because the latter includes the entire history of the covariates while (37) conditions only on current and past $\mathbf{x}_{it}$. Dynamic completeness is natural when $\mathbf{x}_{it}$ contains lagged dependent variables, because it basically means enough lags have been included to capture all of the dynamics. It is often too restrictive in finite distributed lag models such as (30), where (37) would imply

$$\begin{aligned} &E(y_{it}|\mathbf{z}_{it}, y_{i,t-1}\mathbf{z}_{i,t-1}, \ldots, y_{i1}, \mathbf{z}_{i1}, c_i) \\ &\quad = E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i-L}, c_i)\,, \quad t = 1, \ldots, T\,, \end{aligned} \tag{38}$$

which puts strong restrictions on the fully dynamic conditional mean: values y_{ir}, $r \leq t-1$, do not help to predict y_{it} once $(\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots)$ are controlled for. FDLs are of interest even if (38) does not hold. Imposing (37) in FDLs implies that the idiosyncratic errors must be serially uncorrelated, something that is often violated in FDLs.

Dynamic completeness is natural in a model such as

$$y_{it} = \rho y_{i,t-1} + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + c_i + u_{it}\,. \tag{39}$$

Usually – although there are exceptions – (39) is supposed to represent the conditional mean $E(y_{it}|\mathbf{z}_{it}, y_{i,t-1}\mathbf{z}_{i,t-1}, \ldots, y_{i1}, \mathbf{z}_{i1}, c_i)$, and then the issue is whether one lag of y_{it} and $\mathbf{z}_{it}$ suffice to capture the dynamics.

Regardless of what is contained in $\mathbf{x}_{it}$, assumption (37) implies some additional moment conditions that can be used to estimate $\boldsymbol{\beta}$. The extra moment conditions, first proposed in [1] in the context of the AR(1) unobserved effects model, can be written as

$$\begin{aligned} &E[(\Delta y_{i,t-1} - \Delta\mathbf{x}_{i,t-1}\boldsymbol{\beta})'(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}\,, \\ &\qquad\qquad t = 3, \ldots, T\,; \end{aligned} \tag{40}$$

see also [9]. The conditions can be used in conjunction with those in Eq. (33) in a method of moments estimation method. In addition to imposing dynamic completeness, the moment conditions in (40) are nonlinear in parameters, which makes them more difficult to implement than just using (33). Nevertheless, the simulation evidence in [1] for the AR(1) model shows that (40) can help considerably when the coefficient ρ is large.

[7] suggested a different set of restrictions,

$$\mathrm{Cov}(\Delta\mathbf{x}'_{it}, c_i) = 0\,, \quad t = 2, \ldots, T\,. \tag{41}$$

Interestingly, this assumption is very similar in spirit to assumption (23), except that it is in terms of the first difference of the covariates, not the time-demeaned covariates. Condition (41) generates moment conditions in the levels of equation,

$$E\big[\Delta\mathbf{x}'_{it}(y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})\big] = \mathbf{0}\,, \quad t = 2, \ldots, T\,, \tag{42}$$

where α allows for a nonzero mean for c_i. [10] applies these moment conditions, along with the usual conditions in (33), to estimate firm-level production functions. Because of persistence in the data, they find the moments in (33) are not especially informative for estimating the parameters, whereas (42) along with (33) are. Of course, (42) is an extra set of assumptions.

The previous discussion can be applied to the AR(1) model, which has received much attention. In its simplest form the model is

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}\,, \quad t = 1, \ldots, T\,, \tag{43}$$

so that, by convention, the first observation on y is at $t = 0$. The minimal assumptions imposed are

$$E(y_{is}u_{it}) = 0\,, \quad s = 0, \ldots, t-1\,, \quad t = 1, \ldots, T\,, \tag{44}$$

in which case the available instruments at time t are $\mathbf{w}_{it} = (y_{i0}, \ldots, y_{i,t-2})$ in the FD equation

$$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta u_{it}\,, \quad t = 2, \ldots, T\,. \tag{45}$$

Written in terms of the parameters and observed data, the moment conditions are

$$\begin{aligned} &E[y_{is}(\Delta y_{it} - \rho\Delta y_{i,t-1}) = 0\,, \\ &\qquad s = 0, \ldots, t-2\,, \quad t = 2, \ldots, T\,. \end{aligned} \tag{46}$$

[4] proposed pooled IV estimation of the FD equation with the single instrument $y_{i,t-2}$ (in which case all $T-1$ periods can be used) or $\Delta y_{i,t-2}$ (in which case only $T-2$ periods can be used). A better approach is pooled IV where $T-1$ separate reduced forms are estimated for $\Delta y_{i,t-1}$ as a linear function of $(y_{i0}, \ldots, y_{i,t-2})$. The fitted values $\widehat{\Delta y}_{i,t-1}$, can be used as the instruments in (45) in a pooled IV estimation. Of course, standard errors and inference should be made robust to the MA(1) serial correlation in Δu_{it}. [6] suggested full GMM estimation using all of the available instruments $(y_{i0}, \ldots, y_{i,t-2})$, and this estimator uses the conditions in (44) efficiently.

Under the dynamic completeness assumption

$$E(u_{it}|y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}, c_i) = 0\,, \tag{47}$$

the extra moment conditions in [1] become

$$\begin{aligned} &E[(\Delta y_{i,t-1} - \rho\Delta y_{i,t-2})(y_{it} - \rho y_{i,t-1})] = 0\,, \\ &\qquad\qquad t = 3, \ldots, T\,. \end{aligned} \tag{48}$$

[10] noted that if the condition

$$\mathrm{Cov}(\Delta y_{i1}, c_i) = \mathrm{Cov}(y_{i1} - y_{i0}, c_i) = 0 \tag{49}$$

is added to (47) then the combined set of moment conditions becomes

$$E[\Delta y_{i,t-1}(y_{it}-\alpha-\rho y_{i,t-1})]=0\,,\quad t=2,\ldots,T\,,\quad (50)$$

which can be added to the usual moment conditions (46). Conditions (46) and (50) combined are attractive because they are linear in the parameters, and they can produce much more precise estimates than just using (46).

As discussed in [10], condition (49) can be interpreted as a restriction on the initial condition, y_{i0}, and the steady state. When $|\rho|<1$, the steady state of the process is $c_i/(1-\rho)$. Then, it can be shown that (49) holds if the deviation of y_{i0} from its steady state is uncorrelated with c_i. Statistically, this condition becomes more useful as ρ approaches one, but this is when the existence of a steady state is most in doubt. [22] shows theoretically that such restrictions can greatly increase the information about ρ.

Other approaches to dynamic models are based on maximum likelihood estimation. Approaches that condition on the initial condition y_{i0}, suggested by [10,13,15], seem especially attractive. Under normality assumptions, maximum likelihood conditional on y_{i0} is tractable.

If some strictly exogenous variables are added to the AR(1) model, then it is easiest to use IV methods on the FD equation, namely,

$$\Delta y_{it}=\rho\Delta y_{i,t-1}+\Delta\mathbf{z}_{it}\boldsymbol{\gamma}+\Delta u_{it}\,,\qquad t=1,\ldots,T\,.\quad (51)$$

The available instruments (in addition to time period dummies) are $(\mathbf{z}_i, y_{i,t-2},\ldots,y_{i0})$, and the extra conditions (42) can be used, too. If sequentially exogenous variables, say $\mathbf{h}_{it}$, are added, then $(\mathbf{h}_{i,t-1},\ldots,\mathbf{h}_{i1})$ would be added to the list of instruments (and $\Delta\mathbf{h}_{it}$ would appear in the equation).

Unbalanced Panel Data Sets

The previous sections considered estimation of models using balanced panel data sets, where each unit is observed in each time period. Often, especially with data at the individual, family, or firm level, data are missing in some time periods – that is, the panel data set is *unbalanced*. Standard methods, such as fixed effects, can often be applied to produce consistent estimators, and most software packages that have built-in panel data routines typically allow unbalanced panels. However, determining whether applying standard methods to the unbalanced panel produces consistent estimators requires knowing something about the mechanism generating the missing data.

Methods based on removing the unobserved effect warrant special attention, as they allow some nonrandomness in the sample selection. Let $t=1,\ldots,T$ denote the time periods for which data can exist for each unit from the population, and again consider the model

$$y_{it}=\eta_t+\mathbf{x}_{it}\boldsymbol{\beta}+c_i+u_{it}\,,\quad t=1,\ldots,T\,.\quad (52)$$

It is helpful to have, for each i and t, a binary selection variable, s_{it}, equal to one of the data for unit i in time t can be used, and zero otherwise. For concreteness, consider the case where time averages are removed to eliminate c_i, but where the averages necessarily only include the $s_{it}=1$ observations. Let $\ddot{y}_{it}=y_{it}-T_i^{-1}\sum_{r=1}^{T}s_{ir}y_{ir}$ and $\ddot{\mathbf{x}}_{it}=\mathbf{x}_{it}-T_i^{-1}\sum_{r=1}^{T}s_{ir}\mathbf{x}_{ir}$ be the time-demeaned quantities using the observed time periods for unit i, where $T_i=\sum_{t=1}^{T}s_{it}$ is the number of time periods observed for unit i – properly viewed as a random variable. The fixed effects estimator on the unbalanced panel can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{FE}}&=\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\\&\quad\cdot\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}s_{it}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right)\\&=\boldsymbol{\beta}+\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\\&\quad\cdot\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}s_{it}\ddot{\mathbf{x}}_{it}'u_{it}\right)\,.\end{aligned}\quad (53)$$

With fixed T and $N\to\infty$ asymptotics, the key condition for consistency is

$$\sum_{t=1}^{T}E(s_{it}\ddot{\mathbf{x}}_{it}'u_{it})=\mathbf{0}\,.\quad (54)$$

In evaluating (54), it is important to remember that $\ddot{\mathbf{x}}_{it}$ depends on $(\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},s_{i1},\ldots,s_{iT})$, and in a nonlinear way. Therefore, it is not sufficient to assume $(\mathbf{x}_{ir},s_{ir})$ are uncorrelated with u_{it} for all r and t. A condition that is sufficient for (54) is

$$E(u_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},s_{i1},\ldots,s_{iT},c_i)=0\,,\qquad t=1,\ldots,T\,.\quad (55)$$

Importantly, (55) allows arbitrary correlation between the heterogeneity, c_i, and selection, s_{it}, in any time period t. In other words, some units are allowed to be more likely to be in or out of the sample in any time period, and these probabilities can change across t. But (55) rules out some important kinds of sample selection. For example, selection at time t, s_{it}, cannot be correlated with the idiosyncratic error at time t, u_{it}. Further, feedback is not allowed: in affect, like the covariates, selection must be strictly exogenous conditional on c_i.

Testing for no feedback into selection is easy in the context of FE estimation. Under (55), $s_{i,t+1}$ and u_{it} should be uncorrelated. Therefore, $s_{i,t+1}$ can be added to the FE estimation on the unbalanced panel – where the last time period is lost for all observations – and a t test can be used to determine significance. A rejection means (55) is false. Because serial correlation and heteroskedasticity are always a possibility, the t test should be made fully robust.

Contemporaneous selection bias – that is, correlation between s_{it} and u_{it} –is more difficult to test. Chapter 17 in [55] summarizes how to derive tests and corrections by extending the corrections in [28] (so-called "Heckman corrections") to panel data.

First differencing can be used on unbalanced panels, too, although straight first differencing can result in many lost observations: a time period is used only if it is observed along with the previous or next time period. FD is more useful in the case of attrition in panel data, where a unit is observed until it drops out of the sample and never reappears. Then, if a data point is observed at time t, it is also observed at time $t-1$. Differencing can be combined with the approach in [28] to solve bias due to attrition – at least under certain assumptions. See Chap. 17 in [55].

Random effects methods can also be applied with unbalanced panels, but the assumptions under which the RE estimator is consistent are stronger than for FE. In addition to (55), one must assume selection is unrelated to c_i. A natural assumption, that also imposes exogeneity on the covariates with respect to c_i, is

$$E(c_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},s_{i1},\ldots,s_{iT}) = E(c_i)\,. \tag{56}$$

The only case beside randomly determined sample selection where (56) holds is when s_{it} is essentially a function of the observed covariates. Even in this case, (56) requires that the unobserved heterogeneity is mean independent of the observed covariates – as in the typical RE analysis on balanced panel.

Nonlinear Models

Nonlinear panel data models are considerably more difficult to interpret and estimate than linear models. Key issues concern how the unobserved heterogeneity appears in the model and how one accounts for that heterogeneity in summarizing the effects of the explanatory variables on the response. Also, in some cases, conditional independence of the response is used to identify certain parameters and quantities.

Basic Issues and Quantities of Interest

As in the linear case, the setup here is best suited for situations with small T and large N. In particular, the asymptotic analysis underlying the discussion of estimation is with fixed T and $N \to \infty$. Sampling is assumed to be random from the population. Unbalanced panels are generally difficult to deal with because, except in special cases, the unobserved heterogeneity cannot be completely eliminated in obtaining estimating equations. Consequently, methods that model the conditional distribution of the heterogeneity conditional on the entire history of the covariates – as we saw with the Chamberlain–Mundlak approach – are relied on heavily, and such approaches are difficult when data are missing on the covariates for some time periods. Therefore, this section considers only balanced panels. The discussion here takes the response variable, y_{it}, as a scalar for simplicity.

The starting point for nonlinear panel data models with unobserved heterogeneity is the conditional distribution

$$D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)\,, \tag{57}$$

where $\mathbf{c}_i$ is the unobserved heterogeneity for observation i drawn along with the observables. Often there is a particular feature of this distribution, such as $E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$, or a conditional median, that is of primary interest. Even focusing on the conditional mean raises some tricky issues in models where $\mathbf{c}_i$ does not appear in an additive or linear form. To be precise, let $E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t,\mathbf{c})$ be the mean function. If x_{tj} is continuous, then the partial effect can be defined as

$$\theta_j(\mathbf{x}_t,\mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t,\mathbf{c})}{\partial x_{tj}}\,. \tag{58}$$

For discrete (or continuous) variables, (58) can be replaced with discrete changes. Either way, a key question is: How can one account for the unobserved $\mathbf{c}$ in (58)? In order to estimate magnitudes of effects, sensible values of $\mathbf{c}$ need to be plugged into (58), which means knowledge of at least some distributional features of $\mathbf{c}_i$ is needed. For example, suppose $\boldsymbol{\mu}_{\mathbf{c}} = E(\mathbf{c}_i)$ is identified. Then the *partial effect at the average* (PEA),

$$\theta_j(\mathbf{x}_t,\boldsymbol{\mu}_{\mathbf{c}})\,, \tag{59}$$

can be identified if the regression function m_t is identified. Given more information about the distribution of $\mathbf{c}_i$, different quantiles can be inserted into (59), or a certain number of standard deviations from the mean.

An alternative to plugging in specific values for $\mathbf{c}$ is to average the partial effects across the distribution of $\mathbf{c}_i$:

$$\mathrm{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)] , \tag{60}$$

the so-called *average partial effect* (APE). The difference between (59) and (60) can be nontrivial for nonlinear mean functions. The definition in (60) dates back at least to [17], and is closely related to the notion of the *average structural function* (ASF), as introduced in [12]. The ASF is defined as

$$\mathrm{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] . \tag{61}$$

Assuming the derivative passes through the expectation results in (60); computing a discrete change in the ASF always gives the corresponding APE. A useful feature of APEs is that they can be compared across models, where the functional form of the mean or the distribution of the heterogeneity can be different. In particular, APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Average partial effects are not always identified, even when parameters are. Semi-parametric panel data methods that are silent about the distribution of $\mathbf{c}_i$, unconditionally or conditional on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, cannot generally deliver estimates of APEs, essentially by design. Instead, an index structure is usually imposed so that parameters can be consistently estimated. A common setup with scalar heterogeneity is

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c) , \tag{62}$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable. The partial effects are proportional to the parameters:

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c) , \tag{63}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Therefore, if β_j is identified, then so is the sign of the partial effect, and even the relative effects of any two continuous variables: the ratio of partial effects for x_{tj} and x_{th} is β_j/β_h. However, even if $G(\cdot)$ is specified (the common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i; otherwise, the term $E[g(\mathbf{x}_t\boldsymbol{\beta} + c_i)]$ cannot generally be estimated. The probit example below shows how the APEs can be estimated in index models under distributional assumptions for c_i.

The previous discussion holds regardless of the exogeneity assumptions on the covariates. For example, the definition of the APE for a continuous variable holds whether $\mathbf{x}_t$ contains lagged dependent variables or only contemporaneous variables. However, approaches for estimating the parameters and the APEs depend critically on exogeneity assumptions.

Exogeneity Assumptions on the Covariates

As in the case of linear models, it is not nearly enough to simply specify a model for the conditional distribution of interest, $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or some feature of it, in order to estimate parameters and partial effects. This section offers two exogeneity assumptions on the covariates that are more restrictive versions of the linear model assumptions.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) , \tag{64}$$

which means that $\mathbf{x}_{ir}$, $r \neq t$, does not appear in the conditional distribution of y_{it} once $\mathbf{x}_{it}$ and $\mathbf{c}_i$ have been counted for. [17] labeled (64) *strict exogeneity conditional on the unobserved effects* $\mathbf{c}_i$. Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) , \tag{65}$$

which already played a role in linear models. Assumption (64), or its conditional mean version, are less restrictive than if $\mathbf{c}_i$ is not in the conditioning set, as discussed in [17]. Indeed, it is easy to see that, if (64) holds and $D(\mathbf{c}_i|\mathbf{x}_i)$ depends on $\mathbf{x}_i$, then strict exogeneity without conditioning on $\mathbf{c}_i$, $D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$, cannot hold. Unfortunately, both (64) and (65) rule out lagged dependent variables, as well as other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in $\mathbf{x}_{ir}$, $r > t$. Nevertheless, the conditional strict exogeneity assumption underlies the most common estimation methods for nonlinear models.

More natural is *sequential exogeneity conditional on the unobserved effects*, which, in terms of conditional distributions, is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) . \tag{66}$$

Assumption (66) allows for lagged dependent variables and does not restrict feedback. Unfortunately, (66) is substantially more difficult to work with than (64) for general nonlinear models.

Because $\mathbf{x}_{it}$ is conditioned on, neither (64) nor (66) allows for contemporaneous endogeneity of $\mathbf{x}_{it}$ as would arise with measurement error, time-varying omitted variables, or simultaneous equations. This chapter does not treat such cases. See [38] for a recent summary.

Conditional Independence Assumption

The exogeneity conditions stated in Subsect. "Exogeneity Assumptions on the Covariates" generally do not restrict the dependence in the responses, $\{y_{it}\colon t = 1, \ldots, T\}$. Often, a *conditional independence* assumption is explicitly imposed, which can be written generally as

$$D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it} | \mathbf{x}_i, \mathbf{c}_i) . \qquad (67)$$

Equation (67) simply means that, conditional on the entire history $\{\mathbf{x}_{it}\colon t = 1, \ldots, T\}$ and the unobserved heterogeneity $\mathbf{c}_i$, the responses are independent across time. One way to think about (67) is that time-varying unobservables are independent over time. Because (67) conditions on $\mathbf{x}_i$, it is useful only in the context of the strict exogeneity assumption (64). Then, conditional independence can be written as

$$D(y_{i1}, \ldots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i) . \qquad (68)$$

Therefore, under strict exogeneity and conditional independence, the panel data modeling exercise reduces to specifying a model for $D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$. In random effects and correlated RE frameworks, conditional independence can play a critical role in being able to estimate the parameters and the distribution of $\mathbf{c}_i$. As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. Before explaining how that works, the key issue of dependence between the heterogeneity and covariates needs to be addressed.

Assumptions About the Unobserved Heterogeneity

For general nonlinear models, the *random effects assumption* is independence between $\mathbf{c}_i$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$:

$$D(\mathbf{c}_i | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i) . \qquad (69)$$

Assumption (69) is very strong. To illustrate how strong it is, suppose that (69) is combined with a model for the conditional mean, $E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t, \mathbf{c})$. Without any additional assumptions, the average partial effects are nonparametrically identified. In particular, the APEs can be obtained directly from the conditional mean

$$r_t(\mathbf{x}_t) \equiv E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t) . \qquad (70)$$

(The argument is a simple application of the law of iterated expectations; it is discussed in [56]). Nevertheless, (69) is still common in many applications, especially when the explanatory variables of interest do not change over time.

As in the linear case, a *correlated random effects* (CRE) framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but the dependence in restricted in some way. In a parametric setting, a CRE approach involves specifying a distribution for $D(\mathbf{c}_i | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as in [15,17,46], and many subsequent authors; see, for example, [55] and [14]. For many models – see, for example, Subsect. "Binary Response Models" – one can allow $D(\mathbf{c}_i | \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ in a "nonexchangeable" manner, that is, the distribution need not be symmetric on its conditioning arguments. However, allowing nonexchangeability usually comes at the expense of potentially restrictive distributional assumptions, such as homoskedastic normal with a linear conditional mean. For estimating APEs, it is sufficient to assume, along with strict exogeneity,

$$D(\mathbf{c}_i | \mathbf{x}_i) = D(\mathbf{c}_i | \bar{\mathbf{x}}_i) , \qquad (71)$$

without specifying $D(\mathbf{c}_i | \bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. (See, for example, [3,56].) As a practical matter, it makes sense to adopt (71) – or perhaps allow other features of $\{\mathbf{x}_{it}\colon t = 1, \ldots, T\}$ – in a flexible parametric analysis.

Condition (71) still imposes restrictions on $D(\mathbf{c}_i | \mathbf{x}_i)$. Ideally, as in the linear model, one could estimate at least some features of interest without making any assumption about $D(\mathbf{c}_i | \mathbf{x}_i)$. Unfortunately, the scope for allowing unrestricted $D(\mathbf{c}_i | \mathbf{x}_i)$ is limited to special nonlinear models, at least with small T. Allowing $D(\mathbf{c}_i | \mathbf{x}_i)$ to be unspecified is the hallmark of a "fixed effects" analysis, but the label has not been used consistently. Often, fixed effects has been used to describe a situation where the $\mathbf{c}_i$ are treated as parameters to be estimated, along with parameters that do not vary across i. Except in special cases or with large T, estimating the unobserved heterogeneity is prone to an *incidental parameters problem*. Namely, using a fixed T, $N \to \infty$ framework, one cannot get consistent estimators of the $\mathbf{c}_i$, and the inconsistency in, say, $\hat{\mathbf{c}}_i$, generally transmits itself to the parameters that do not vary with i. The incidental parameters problem does not arise in estimating the coefficients $\boldsymbol{\beta}$ in a linear model because the estimator obtained by treating the c_i as parameters to estimate is equivalent to pooled OLS on the time-demeaned data –

that is, the fixed effects estimator can be obtained by eliminating the c_i using the within transformation or estimating the c_i along with $\boldsymbol{\beta}$. This occurrence is rare in nonlinear models. Section "Future Directions" further discusses this issue, as there is much ongoing research that attempts to reduce the asymptotic bias in nonlinear models.

The "fixed effects" label has also been applied to settings where the $\mathbf{c}_i$ are not treated as parameters to estimate; rather, the $\mathbf{c}_i$ can be eliminated by conditioning on a sufficient statistic. Let $\mathbf{w}_i$ be a function of the observed data, $(\mathbf{x}_i, \mathbf{y}_i)$, such that

$$D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{w}_i) = D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{w}_i) . \quad (72)$$

Then, provided the latter conditional distribution depends on the parameters of interest, and can be derived or approximated from the original specification of $D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i)$, maximum likelihood methods can be used. Such an approach is also called *conditional maximum likelihood estimation* (CMLE), where "conditional" refers to conditioning on a function of $\mathbf{y}_i$. (In traditional treatments of MLE, conditioning on so-called "exogenous" variables is usually implicit.) In most cases where the CMLE approach applies, the conditional independence assumption (67) is maintained, although one conditional MLE is known to have robustness properties: the so-called "fixed effects" Poisson estimator (see [53]).

Maximum Likelihood Estimation and Partial MLE

There are two common approaches to estimating the parameters in nonlinear, unobserved effects panel data models when the explanatory variables are strictly exogenous. (A third approach, generalized method of moments, is available in special cases but is not treated here. See, for example, Chap. 19 in [55].) The first approach is full maximum likelihood (conditional on the entire history of covariates). Most commonly, full MLE is applied under the conditional independence assumption, although sometimes models are used that explicitly allow dependence in $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, \mathbf{c}_i)$. Assuming strict exogeneity, conditional independence, a model for the density of y_{it} given $(\mathbf{x}_{it}, \mathbf{c}_i)$ (say, $f_t(y_t|\mathbf{x}_t, \mathbf{c}; \theta)$), and a model for the density of $\mathbf{c}_i$ given $\mathbf{x}_i$ (say, $h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})$), the log likelihood for random draw i from the cross section is

$$\log\left\{\left[\int \prod_{t=1}^{T} f_t(y_{it}|\mathbf{x}_{it}, \mathbf{c}; \theta)\right] h(\mathbf{c}|\mathbf{x}_i; \boldsymbol{\delta})\mathrm{d}\mathbf{c}\right\} . \quad (73)$$

This log-likelihood function "integrates out" the unobserved heterogeneity to obtain the joint density of $(y_{i1}, \ldots, y_{iT})$ conditional on $\mathbf{x}_i$. In the most commonly applied models, including logit, probit, Tobit, and various count models (such as the Poisson model), the log likelihood in (73) identifies all of the parameters. Computation can be expensive but is typically tractable. The main methodological drawback to the full MLE approach is that it is not robust to violations of the conditional independence assumption, except for the linear model where normal conditional distributions are used for y_{it} and c_i.

The *partial* MLE ignores temporal dependence in the responses when estimating the parameters – at least when the parameters are identified. In particular, obtain the density of y_{it} given $\mathbf{x}_i$ by integrating the marginal density for y_{it} against the density for the heterogeneity:

$$g_t(y_t|\mathbf{x}; \theta, \boldsymbol{\delta}) = \int f_t(y_t|\mathbf{x}_t, \mathbf{c}; \theta)h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})\,\mathrm{d}\mathbf{c} . \quad (74)$$

The *partial* MLE (PMLE) (or pooled MLE) uses, for each i, the partial log likelihood

$$\sum_{t=1}^{T} \log[g_t(y_{it}|\mathbf{x}_i; \theta, \boldsymbol{\delta}) . \quad (75)$$

Because the partial MLE ignores the serial dependence caused by the presence of $\mathbf{c}_i$, it is essentially never efficient. But in leading cases, such as probit, Tobit, and Poisson models, $g_t(y_t|\mathbf{x}; \theta, \boldsymbol{\delta})$ has a simple form when $h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})$ is chosen judiciously. Further, the PMLE is fully robust to violations of (67). Inference is complicated by the neglected serial dependence, but an appropriate adjustment to the asymptotic variance is easily obtained; see Chap. 13 in [55].

One complication with PMLE is that in the cases where it leads to a simple analysis (probit, ordered probit, and Tobit, to name a few), not all of the parameters in θ and $\boldsymbol{\delta}$ are separately identified. The conditional independence assumption and the use of full MLE serves to identify all parameters. Fortunately, the PMLE does identify the parameters that index the average partial effects, a claim that will be verified for the probit model in Subsect. "Binary Response Models".

Dynamic Models

General models with only sequentially exogenous variables are difficult to estimate. [8] considered binary response models and [54] suggested a general strategy that requires modeling the dynamic distribution of the variables that are not strictly exogenous.

Much more is known about the specific case where the model contains lagged dependent variables along with strictly exogenous variables. The starting point is a model for the dynamic distribution,

$$D(y_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i)\,, \qquad t = 1, \ldots, T\,, \tag{76}$$

where $\mathbf{z}_{it}$ are variables strictly exogenous (conditional on $\mathbf{c}_i$) in the sense that

$$\begin{aligned} &D(y_{it}|\mathbf{z}_i, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i) \\ &\quad = D(\mathbf{y}_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i)\,, \end{aligned} \tag{77}$$

where $\mathbf{z}_i$ is the entire history $\{\mathbf{z}_{it}\colon t = 1, \ldots, T\}$.

In the leading case, (76) depends only on $(\mathbf{z}_{it}, y_{i,t-1}, \mathbf{c}_i)$ (although putting lags of strictly exogenous variables only slightly changes the notation). Let $f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta)$ denote a model for the conditional density, which depends on parameters θ. The joint density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta)\,. \tag{78}$$

The problem with using (78) for estimation is that, when it is turned into a log likelihood by plugging in the "data", $\mathbf{c}_i$ must be inserted. Plus, the log likelihood depends on the initial condition, y_{i0}. Several approaches have been suggested to address these problems: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (which results in an incidental parameters problem). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for c_i. (This approach is available for very limited situations, and other restrictions are needed. And, generally, one cannot estimate average partial effects.) (iii) Find, or, more practically, approximate $D(y_{i0}|\mathbf{c}_i, z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. Integrating out $\mathbf{c}_i$ gives the density for $D(y_{i0}, y_{i1}, \ldots, y_{iT}|\mathbf{z}_i)$, which can be used in an MLE analysis (conditional on $\mathbf{z}_i$), (iv) Model $D(\mathbf{c}_i|y_{i0}, \mathbf{z}_i)$. Then, integrate out $\mathbf{c}_i$ conditional on $(y_{i0}, \mathbf{z}_i)$ to obtain the density for $D(y_{i1}, \ldots, y_{iT}|y_{i0}, \mathbf{z}_i)$. Now, MLE is conditional on $(y_{i0}, \mathbf{z}_i)$. As shown by [57], in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|y_0, \mathbf{z})$ that mixes with the density $f(y_1, \ldots, y_T|y_0, \mathbf{z}, \mathbf{c})$ to produce a log-likelihood that is in a common family and programmed in standard software packages.

If $m_t(\mathbf{x}_t, \mathbf{c}, \theta)$ is the mean function $E(y_t|\mathbf{x}_t, \mathbf{c})$, with $\mathbf{x}_t = (\mathbf{z}_t, y_{t-1})$, then APEs are easy to obtain. The average structural function is

$$\begin{aligned} \mathrm{ASF}(\mathbf{x}_t) &= E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i, \theta)] \\ &= E\left\{\left[\int m_t(\mathbf{x}_t, \mathbf{c}, \theta) h(\mathbf{c}|y_{i0}, \mathbf{z}_i, \boldsymbol{\gamma}) \mathrm{d}\mathbf{c}\right] | y_{i0}, \mathbf{z}_i\right\}. \end{aligned} \tag{79}$$

The term inside the brackets, say $r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \theta, \boldsymbol{\gamma})$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, or they can be simulated. A consistent estimator of the ASF is obtained by averaging out $(y_{i0}, \mathbf{z}_i)$:

$$\widehat{\mathrm{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{t=1}^{T} r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \hat{\theta}, \hat{\boldsymbol{\gamma}})\,. \tag{80}$$

Partial derivatives and differences with respect to elements of $\mathbf{x}_t$ (which, remember, includes functions of y_{t-1}) can be computed. With large N and small T, the panel data bootstrap – where resampling is carried out in the cross section so that every time period is kept when a unit i is resampled – can be used for standard errors and inference. The properties of the nonparametric bootstrap are standard in this setting because the resampling is carried out in the cross section.

Binary Response Models

Unobserved effects models – static and dynamic – have been estimated for various kinds of response variables, including binary responses, ordered responses, count data, and corner solutions. Most of the issues outlined above can be illustrated by binary response models, which is the topic of this subsection.

The standard specification for the unobserved effects (UE) probit model is

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)\,, \quad t = 1, \ldots, T\,, \tag{81}$$

where $\mathbf{x}_{it}$ does not contain an overall intercept but would usually include time dummies, and c_i is the scalar heterogeneity. Without further assumptions, neither $\boldsymbol{\beta}$ nor the APEs are identified. The traditional RE probit model imposes a strong set of assumptions: strict exogeneity, conditional independence, and independence between c_i and $\mathbf{x}_i$ with $c_i \sim \mathrm{Normal}(\mu_c, \sigma_c^2)$. Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE (conditional on $\mathbf{x}_i$).

Under the strict exogeneity assumption (64), a correlated random effects version of the model is obtained from the Chamberlain–Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim \mathrm{Normal}(0, \sigma_a^2)\,. \tag{82}$$

The less restrictive version $c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1}\boldsymbol{\xi}_1 + \cdots + \mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i$ can be used, but the time average conserves on degrees of freedom.

As an example, suppose that y_{it} is a binary variable indicating whether firm i in year t was awarded at least one patent, and the key explanatory variable in $\mathbf{x}_{it}$ is current and past spending on research and development (R&D). It makes sense that R&D spending is correlated, at least on average, with unobserved firm heterogeneity, and so a correlated random effects model seems natural. Unfortunately, the strict exogeneity assumption might be problematical: it could be that being awarded a patent in year t might affect future values of spending on R&D. Most studies assume this is not the case, but one should be aware that, as in the linear case, the strict exogeneity assumption imposes restrictions on economic behavior.

When the conditional independence assumption (67) is added to (81), strict exogeneity, and (82), all parameters in (81) and (82) are identified (assuming that all elements of $\mathbf{x}_{it}$ are time-varying) and the parameters can be efficiently estimated by maximum likelihood (conditional on $\mathbf{x}_i$). Afterwards, the mean of c_i can be consistently estimated as $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 = \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Because a_i is normally distributed, c_i is not normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. A normal approximation for $D(c_i)$ gets better as T gets large. In any case, the estimated mean and standard deviation can be used to plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$ or $\hat{\mu}_c \pm 2\hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated by

$$\widehat{\mathrm{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{83}$$

where the "a" subscript means that a coefficient has been divided by $(1+\hat{\sigma}_a^2)^{1/2}$, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\mathrm{ASF}}(\mathbf{x}_t)$ with respect to elements of $\mathbf{x}_t$ can be compared with fixed effects estimates from a linear model. Often, to obtain a single scale factor, a further averaging across $\mathbf{x}_{it}$ is done. The APEs computed from such averaging can be compared to linear FE estimates.

The CRE probit model is an example of a model where the APEs are identified without the conditional independence assumption. Without (67) – or any restriction on the joint distribution – it can still be shown that

$$P(y_{it}=1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)\,, \tag{84}$$

which means a number of estimation approaches identify the scaled coefficients $\boldsymbol{\beta}_a$, ψ_a, and $\boldsymbol{\xi}_a$. The estimates of these scaled coefficients can be inserted directly into (83). The unscaled parameters and σ_a^2 are not separately identified, but in most cases this is a small price to pay for relaxing the conditional independence assumption. Note that for determining directions of effects and relative effects, $\boldsymbol{\beta}_a$ is just as useful as $\boldsymbol{\beta}$. Plus, it is $\boldsymbol{\beta}_a$ that appears in the APEs. The partial effects at the mean value of c_i are not identified.

Using pooled probit can be inefficient for estimating the scaled parameters. Full MLE, with a specified correlation matrix for the $T \times 1$ vector $\mathbf{u}_i$, is possible in principle but difficult in practice. An alternative approach, the *generalized estimating equations* (GEE) approach, can be more efficient than pooled probit but just as robust in that only (84) is needed for consistency. See [38] for a summary of how GEE – which is essentially the same as multivariate weighted nonlinear least squares – applies to the CRE probit model.

A simple test of the strict exogeneity assumption is to add selected elements of $\mathbf{x}_{i,t+1}$, say $\mathbf{w}_{i,t+1}$, to the model and computing a test of joint significance. Unless the full MLE is used, the test should be made robust to serial dependence of unknown form. For example, as a test of strict exogeneity of R&D spending when y_{it} is a patent indicator, one can just include next year's value of R&D spending and compute a t test. In carrying out the test, the last time period is lost for all firms.

Because there is nothing sacred about the standard model (81) under (82) – indeed, these assumptions are potentially quite restrictive – it is natural to pursue other models and assumptions. Even with (81) as the starting point, and under strict exogeneity, there are no known ways of identifying parameters or partial effects without restricting $D(c_i|\mathbf{x}_i)$. Nevertheless, as mentioned in Subsect. "Assumptions About the Unobserved Heterogeneity", there are nonparametric restrictions on $D(c_i|\mathbf{x}_i)$ that do identify the APEs under strict exogeneity – even if (81) is dropped entirely. As shown in [3], the restriction $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$ identifies the APEs. While fully nonparametric methods can be used, some simple strategies are evident. For example, because the APEs can be obtained from $D(y_{it}|\mathbf{x}_{it},\bar{\mathbf{x}}_i)$, it makes sense to apply flexible parametric models directly to this distribution – without worrying about the original models for $D(y_{it}|\mathbf{x}_{it},c_i)$ and $D(c_i|\mathbf{x}_i)$.

As an example of this approach, a flexible parametric model, such as

$$\begin{aligned} &P(y_{it}=1|\mathbf{x}_{it},\bar{\mathbf{x}}_i) \\ &\quad = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}]\,, \end{aligned} \tag{85}$$

might provide a reasonable approximation. The average structural function is estimated as

$$\widehat{\mathrm{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi[\hat{\theta}_t+\mathbf{x}_t\hat{\boldsymbol{\beta}}+\bar{\mathbf{x}}_i\hat{\boldsymbol{\gamma}}+(\bar{\mathbf{x}}_i\otimes\bar{\mathbf{x}}_i)\hat{\boldsymbol{\delta}}+(\mathbf{x}_t\otimes\bar{\mathbf{x}}_i)\hat{\boldsymbol{\eta}}]\,, \tag{86}$$

where the estimates can come from pooled MLE, GEE, or a method of moments procedure. The point is that extensions of the basic probit model such as (85) can provide considerable flexibility and deliver good estimators of the APEs. The drawback is that one has to be willing to abandon standard underlying models for $P(y_{it} = 1|\mathbf{x}_{it}, c_i)$ and $D(c_i|\mathbf{x}_i)$; in fact, it seems very difficult to characterize models for these two features that would lead to an expression such as (85).

An alternative model for the response probability is the logit model

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)\,, \tag{87}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$. In cross section applications, researchers often find few practical differences between (81) and (87). But when unobserved heterogeneity is added in a panel data context, the logit formulation has an advantage: under the conditional independence assumption (and strict exogeneity), the parameters $\boldsymbol{\beta}$ can be consistently estimated, with a $\sqrt{N}$-asymptotic normal distribution, without restricting $D(c_i|\mathbf{x}_i)$. The method works by conditioning on the number of "successes" for each unit, that is, $n_i = \sum_{t=1}^{T} y_{it}$. [17] shows that $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i, n_i) = D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, n_i)$, and the latter depends on $\boldsymbol{\beta}$ (at least when all elements of $\mathbf{x}_{it}$ are time varying). The conditional MLE – sometimes called the "fixed effects logit" estimator – is asymptotically efficient in the class of estimators putting no assumptions on $D(c_i|\mathbf{x}_i)$. While this feature of the logit CMLE is attractive, the method has two drawbacks. First, it does not appear to be robust to violations of the conditional independence assumption, and little is known about the practical effects of serial dependence in $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i)$. Second, and perhaps more importantly, because $D(c_i|\mathbf{x}_i)$ and $D(c_i)$ are not restricted, it is not clear how one estimates magnitudes of the effects of the covariates on the response probability. The logit CMLE is intended to estimate the parameters, which means the effects of the covariates on the log-odds ratio, $\log\{[P(y_{it} = 1|\mathbf{x}_{it}, c_i)]/[1 - P(y_{it} = 1|\mathbf{x}_{it}, c_i)]\} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i$, can be estimated. But the magnitudes of the effects of covariates on the response probability are not available. Therefore, there are tradeoffs when choosing between CRE probit and "fixed effects" logit: the CRE probit identifies average partial effects with or without the conditional independence assumptions, at the cost of specifying $D(c_i|\mathbf{x}_i)$, while the FE logit estimates parameters without specifying $D(c_i|\mathbf{x}_i)$, but requires conditional independence and still does not deliver estimates of partial effects. As often is the case in econometrics, there are tradeoffs between assumptions between the logit and probit approaches, and also tradeoffs. See [38] for further discussion.

Estimation of parameters and APEs is more difficult in simple dynamic probit models. Consider

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i)\,, \tag{88}$$

which assumes first-order dynamics and strict exogeneity of $\{\mathbf{z}_{it}\colon t = 1, \ldots, T\}$. Treating the c_i as parameters to estimate causes inconsistency in $\boldsymbol{\delta}$ and ρ because of the incidental parameters problem. A simple analysis is available under the assumption

$$c_i|y_{i0}, \mathbf{z}_i \sim \mathrm{Normal}(\psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2)\,. \tag{89}$$

Then,

$$\begin{aligned} &P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) \\ &\quad = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i)\,, \end{aligned} \tag{90}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$. Because a_i is independent of $(y_{i0}, \mathbf{z}_i)$, it turns out that standard random effects probit software can be used, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. All parameters, including σ_a^2, are consistently estimated, and the ASF is estimated by averaging out $(y_{i0}, \mathbf{z}_i)$:

$$\widehat{\mathrm{ASF}}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a)\,, \tag{91}$$

where the coefficients are multiplied by $(1+\hat{\sigma}_a^2)^{-1/2}$. APEs are gotten, as usual, by taking differences or derivatives with respect to elements of $(\mathbf{z}_t, y_{t-1})$. Both (88) and the model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $\mathrm{Var}(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic). See [57] for further discussion.

Similar analyses hold for other nonlinear models, although the particulars differ. For count data, maximum likelihood methods are available – based on correlated random effects or conditioning on a sufficient statistic. In this case, the CMLE based on the Poisson distribution has very satisfying robustness properties, requiring only the conditional mean in the unobserved effects model to be correctly specified along with strict exogeneity (Conditional independence is not needed). These and dynamic count models are discussed in Chap. 19 in [55,57].

Correlated random effects Tobit models are specified and estimated in a manner very similar to CRE probit models; see Chap. 16 in [55]. Unfortunately, there are no known conditional MLEs that eliminate the unobserved heterogeneity in Tobit models. Nevertheless, [33,34] show how the parameters in models for corner solutions can be estimated without distributional assumptions on $D(c_i|\mathbf{x}_i)$. Such methods do place exchangeability restrictions on $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i)$, but they are not as strong as conditional independence with identical distributions.

Future Directions

Research in panel data methods continues unabated. Dynamic linear models are a subject of ongoing interest. The problem of feedback in linear models when the covariates are persistent – and the weak instrument problem that it entails – is important for panels with small T. For example, with firm-level panel data, the number of time periods is typically small and inputs into a production function would often be well-approximated as random walks with perhaps additional short-term dependence. The estimators described in Sect. "Sequentially Exogenous Regressors and Dynamic Models" that impose additional assumptions should be studied when those assumptions fail. Perhaps the lower variance of the estimators from the misspecified model is worth the additional bias.

Models with random coefficients, especially when those random coefficients are on non-strictly exogenous variables (such as lagged dependent variables), have received some attention, but many of the proposed solutions require large T. (See, for example, [49,50]). An alternative approach is flexible MLE, as in [57], where one models the distribution of heterogeneity conditional on the initial condition and the history of covariates. See [19] for any application to dynamic product choice.

When T is large enough so that it makes sense to use large-sample approximations with large T, as well as large N, one must make explicit assumptions about the time series dependence in the data. Such frameworks are sensible for modeling large geographical units, such as states, provinces, or countries, where long stretches of time are observed. The same estimators that are attractive for the fixed T case, particularly fixed effects, can have good properties when T grows with N, but the properties depend on whether unit-specific effects, time-specific effects, or both are included. The rates at which T and N are assumed to grow also affect the large-sample approximations. See [52] for a survey of linear model methods with T and N are both assumed to grow in the asymptotic analysis. A recent study that considers estimation when the data have unit roots is [44]. Unlike the fixed T case, a unified theory for linear models, let alone nonlinear models, remains elusive when T grows with N and is an important area for future research.

In the models surveyed here, a single coefficient is assumed for the unobserved heterogeneity, whereas the effect might change over time. In the linear model, the additive c_i can be replaced with $\psi_t c_i$ (with $\psi_1 = 1$ as a normalization). For example, the return to unobserved managerial talent in a firm production function can change over time. Conditions under which ignoring the time-varying loads, ψ_t, and using the usual fixed effects estimator, consistently estimates the coefficients on $\mathbf{x}_{it}$ are given in [47]. But one can also estimate the ψ_t along with $\boldsymbol{\beta}$ using method of moments frameworks. Examples are [2,32]. An area for future research is to allow heterogeneous slopes on observed covariates along with time-varying loads on the unobserved heterogeneity. Allowing for time-varying loads and heterogeneous slopes in nonlinear models can allow for significant flexibility, but only parametric approaches to estimation have been studied.

There is considerable interest in estimating production functions using proxy variables, such as investment, for time-varying, unobserved productivity. The pioneering work is [48]; see also [42]. Estimation in this case does not rely on differencing or time-demeaning to remove unobserved heterogeneity, and so the estimates can be considerably more precise than the FE or FD estimators. But the assumption that a deterministic function of investment can proxy for unobserved productivity is strong. [11] provides an analysis that explicitly allows for unobserved heterogeneity and non-strictly exogenous inputs using the methods described in Sect. "Sequentially Exogenous Regressors and Dynamic Models". An interesting challenge for future researchers is to unify the two approaches to exploit the attractive features of each.

The parametric correlated random effects approach for both static and dynamic nonlinear models is now fairly well understood in the balanced case. Much less attention has been paid to the unbalanced case, and missing data, especially for fully dynamic models, is a serious challenge. [57] discusses the assumptions under which using a balanced subset produces consistent estimates.

Identification of average partial effects (equivalently, the average structural function) has recently received the attention that it deserves, although little is known about how robust are the estimated APEs under various misspecifications of parametric models. One might hope that using flexible models for nonlinear responses might provide good approximations, but evidence on this issue is lacking.

As mentioned earlier, recent research in [3] has shown how to identify and estimate partial effects without making parametric assumptions about $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ or $D(\mathbf{c}_i|\mathbf{x}_i)$. The setup in [3] allows for $D(\mathbf{c}_i|\mathbf{x}_i)$ to depend on $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ in an exchangeable way. The simplest case is the one given in (71), $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$. Under (71) and the strict exogeneity assumption $E(y_{it}|\mathbf{x}_i, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, the average structural function is identified as

$$\mathrm{ASF}_t(\mathbf{x}_t) = E_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)] \,, \tag{92}$$

where $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. Because $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ can be estimated very generally – even using nonparametric regression of y_{it} on $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ for each t – the average partial effects can be estimated without any parametric assumptions. Research in [3] shows how $D(\mathbf{c}_i|\mathbf{x}_i)$ can depend on other exchangeable functions of $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, such as sample variances and covariances. As discussed in [38], nonexchangeable functions, such as trends and growth rates, can be accommodated, provided these functions are known. For example, for each i, let $(\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$ be the vectors of intercepts and slopes from the regression $\mathbf{x}_{it}$ on $1, t$, $t = 1, \ldots, T$. Then, an extension of (71) is $D(c_i|\mathbf{x}_i) = D(c_i|\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$. It appears these kinds of assumptions have not yet been applied, but they are a fertile area for future research because they extend the typical CRE setup.

Future research on nonlinear models will likely consider the issue of the kinds of partial effects that are of most interest. [3] studies identification and estimation of the *local average response* (LAR). The LAR at $\mathbf{x}_t$ for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t) \,, \tag{93}$$

where $m_t(\mathbf{x}_t, \mathbf{c})$ is the conditional mean of the response and $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $D(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a "local" partial effect because it averages out the heterogeneity for the slice of the population described by the vector of observed covariates, $\mathbf{x}_t$. The APE averages out over the entire distribution of $\mathbf{c}_i$, and therefore can be called a "global average response". See also [21]. The results in [3] include general identification results for the LAR, and future empirical researchers using nonlinear panel data models may find the local nature of the LAR more appealing (although more difficult to estimate) than APEs.

A different branch of the panel data literature has studied identification of coefficients or, more often, scaled coefficients, in nonlinear models. For example, [35] shows how to estimate $\boldsymbol{\beta}$ in the model

$$y_{it} = 1[w_{it} + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0] \tag{94}$$

without distributional assumptions on the composite error, $c_i + u_{it}$. In this model, w_{it} is a special continuous explanatory variable (which need not be time varying). Because its coefficient is normalized to unity, w_{it} necessarily affects the response, y_{it}. More importantly, w_{it} is assumed to satisfy the distributional restriction $D(c_i + u_{it}|w_{it}, \mathbf{x}_{it}, \mathbf{z}_i) = D(c_i + u_{it}|\mathbf{x}_{it}, \mathbf{z}_i)$, which is a conditional independence assumption. The vector $\mathbf{z}_i$ is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, $\mathbf{z}_i$ could be functions of variables determined prior to the earliest time period). The most likely scenario where the framework in [35] applies is when w_{it} is randomized and therefore independent of the entire vector $(\mathbf{x}_{it}, \mathbf{z}_i, c_i + u_{it})$. The key condition seems unlikely to hold if w_{it} is related to past outcomes on y_{it}. The estimator of $\boldsymbol{\beta}$ derived in [35] is $\sqrt{N}$-asymptotically normal, and fairly easy to compute; it requires estimation of the density of w_{it} given $(\mathbf{x}_{it}, \mathbf{z}_i)$ and then a simple IV estimation. Essentially by construction, estimation of partial effects on the response probability is not possible.

Recently, [36] shows how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model in Eq. (85) under the strict exogeneity assumption on $\{\mathbf{z}_{it}\colon t = 1, \ldots, T\}$. A further assumption is that c_i and $\mathbf{z}_i$ are independent. By putting restrictions on $D(c_i)$ –which nevertheless allow flexibility – [36] explains how to estimate bounds for the unknown ρ. The bounds allow one to determine how much information are in the data when few assumptions are made. Similar calculations can be made for average partial effects, so that the size of so-called state dependence – the difference between $E_{c_i}[\Phi(\mathbf{z}_t\boldsymbol{\delta} + \rho + c_i) - \Phi(\mathbf{z}_t\boldsymbol{\delta} + c_i)]$ – can be bounded.

Because CRE methods require some restriction on the distribution of heterogeneity, and estimation of scaled coefficients leaves partial effects unidentified, the theoretical literature has returned to the properties of parameter estimates and partial effects when the heterogeneity is treated as unit-specific parameters to estimate. Recent work has focused on adjusting the "fixed effects" estimates (of the common population parameters) so that they have reduced bias.

An emerging question is whether the average partial effects might be estimated well even though the parameters themselves are biased. In other words, suppose that for a nonlinear model one obtains $\{\hat{\theta}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \ldots, \hat{\mathbf{c}}_N\}$, typically by maximizing a pooled log-likelihood function across all i and t. If $m_t(\mathbf{x}_t, \mathbf{c}, \theta,) = E(y_t|\mathbf{x}_t, \mathbf{c})$ is the conditional mean function, the average partial effects can be estimated as

$$N^{-1} \sum_{i=1}^{N} \frac{\partial m_t(\mathbf{x}_t, \hat{\mathbf{c}}_i, \hat{\theta})}{\partial x_{tj}} \,. \qquad (95)$$

In the unobserved effects probit model, (95) becomes

$$N^{-1} \sum_{i=1}^{N} \hat{\beta}_j \phi(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{c}_i) \,. \qquad (96)$$

[20] studied the properties of (96) with strictly exogenous regressors under conditional independence, assuming that the covariates are weakly dependent over time. Interestingly, the bias in (96) is of order T^{-2} when there is no heterogeneity, which suggests that estimating the unobserved effects might not be especially harmful when the amount of heterogeneity is small. Unfortunately, these findings do not carry over to models with time heterogeneity or lagged dependent variables. While bias corrections are available, they are difficult to implement.

[24] proposes both jackknife and analytical bias corrections and show that they work well for the probit case. Generally, the jackknife procedure to remove the bias in $\hat{\theta}$ is simple but can be computationally intensive. The idea is this. The estimator based on T time periods has probability limit (as $N \to \infty$) that can be written as

$$\theta_T = \theta + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \qquad (97)$$

for vectors $\mathbf{b}_1$ and $\mathbf{b}_2$. Now, let $\hat{\theta}_{(t)}$ denote the estimator that drops time period t. Then, assuming stability across t, it can be shown that the jackknife estimator,

$$\tilde{\theta} = T\hat{\theta} - (T-1)T^{-1} \sum_{t=1}^{T} \hat{\theta}_{(t)} \qquad (98)$$

has asymptotic bias of $\tilde{\theta}$ on the order of T^{-2}.

Unfortunately, there are currently some practical limitations to the jackknife procedure, as well as to the analytical corrections derived in [24]. First, aggregate time effects are not allowed, and they would be very difficult to include because the analysis is with $T \to \infty$. (In other words, time effects would introduce an incidental parameters problem in the time dimension, in addition to the incidental parameters problem in the cross section). Plus, heterogeneity in the distribution of the response y_{it} across t changes the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ when a time period is dropped, and so the adjustment in (98) does not remove the bias terms. Second, [24] assumes independence across t conditional on c_i. It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, even without time heterogeneity, the jackknife does not apply to dynamic models; see [23].

Another area that has seen a resurgence is so-called pseudo panel data, as initially exposited in [18]. A pseudo-panel data set is constructed from repeated cross sections across time, where the units appearing in each cross section are not repeated (or, if they are, it is a coincidence and is ignored). If there is a natural grouping of the cross-sectional units – for example, for individuals, birth year cohorts – one can create a pseudo-panel data set by constructing group or cohort averages in each time period. With relatively few cohorts and large cross sections, one can identify pseudo panels in the context of minimum distance estimation. With a large number of groups, a different large-sample analysis might be warranted. A recent contribution is [39] and [38] includes a recent survey. Open questions include the most efficient way to use the full set of restrictions in the underlying individual-level model.

As mentioned earlier, this chapter did not consider panel data model with explanatory variables that are endogenous in the sense that they are correlated with time-varying unobservables. For linear models, the usual fixed effects and first differencing transformations can be combined with instrumental variables methods. In nonlinear models, the Chamberlain–Mundlak approach can be combined with so-called "control function" methods, provided the endogenous explanatory variables are continuous. [38] includes a discussion of some recent advances for complicated models such as multinomial response models; see also [51]. Generally, structural estimation in discrete response models with unobserved heterogeneity and endogenous explanatory variables is an area of great interest.

Bibliography

Primary Literature

1. Ahn SC, Schmidt P (1995) Efficient estimation of models for dynamic panel data. J Econom 68:5–27
2. Ahn SC, Lee YH, Schmidt P (2001) GMM estimation of linear panel data models with time-varying individual effects. J Econom 101:219–255
3. Altonji JG, Matzkin RL (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73:1053–1102
4. Anderson TW, Hsiao C (1982) Formulation and estimation of dynamic models using panel data. J Econom 18:47–82

5. Arellano M (1993) On the testing of correlated effects with panel data. J Econom 59:87–97
6. Arellano M, Bond SR (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev Econ Stud 58:277–297
7. Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error components models. J Econom 68:29–51
8. Arellano M, Carrasco R (2003) Binary choice panel data models with predetermined variables. J Econom 115:125–157
9. Arellano M, Honoré B (2001) Panel data models: Some recent developments. In: Heckman JJ, Leamer E (eds) Handbook of econometrics, vol 5. North Holland, Amsterdam, pp 3229–3296
10. Blundell R, Bond SR (1998) Initial conditions and moment restrictions in dynamic panel data models. J Econom 87:115–143
11. Blundell R, Bond SR (2000) GMM estimation with persistent panel data: An application to production functions. Econom Rev 19:321–340
12. Blundell R, Powell JL (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in economics and econonometrics: Theory and applications, 8th World Congress, vol 2. Cambridge University Press, Cambridge, pp 312–357
13. Blundell R, Smith RJ (1991) Initial conditions and efficient estimation in dynamic panel data models – an application to company investment behaviours. Ann d'écon stat 20–21:109–124
14. Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and applications. Cambridge University Press, Cambridge
15. Chamberlain G (1980) Analysis of covariance with qualitative data. Rev Econ Stud 47:225–238
16. Chamberlain G (1982) Multivariate regression models for panel data. J Econom 1:5–46
17. Chamberlain G (1984) Panel data. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics, vol 2. North Holland, Amsterdam, pp 1248–1318
18. Deaton A (1985) Panel Data from time series of cross-sections. J Econom 30:109–126
19. Erdem T, Sun B (2001) Testing for choice dynamics in panel data. J Bus Econ Stat 19:142–152
20. Fernández-Val I (2007) Fixed effects estimation of structural parameters and marginal effects in panel probit models. Mimeo. Boston University, Department of Economics
21. Florens JP, Heckman JJ, Meghir C, Vytlacil E (2007) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Mimeo. Columbia University, Department of Economics
22. Hahn J (1999) How informative is the initial condition in the dynamic panel model with fixed effects? J Econom 93:309–326
23. Hahn J, Kuersteiner G (2002) Asymptotically unbiased inference for a dynamic panel model with fixed effects when both *n* and *T* are large. Econometrica 70:1639–1657
24. Hahn J, Newey WK (2004) Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72:1295–1319
25. Hausman JA (1978) Specification tests in econometrics. Econometrica 46:1251–1271
26. Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. Econometrica 49:1377–1398
27. Hayashi F (2000) Econometrics. Princeton University Press, Princeton
28. Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 5:475–492
29. Heckman JJ (1981) Statistical models for discrete panel data. In: Manski CF, McFadden DL (eds) Structural analysis of discrete data and econometric applications. MIT Press, Cambridge, pp 114–178
30. Heckman JJ (1981) The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process. In: Manski CF, McFadden DL (eds) Structural analysis of discrete data and econometric applications. MIT Press, Cambridge, pp 179–195
31. Hoch I (1962) Estimation of production function parameters combining time-series and cross-section data. Econometrica 30:34–53
32. Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. Econometrica 56:1371–1395
33. Honoré BE (1992) Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. Econometrica 60:533–565
34. Honoré BE, Hu L (2004) Estimation of cross sectional and panel data censored regression models with endogeneity. J Econom 122:293–316
35. Honoré BE, Lewbel A (2002) Semiparametric binary choice panel data models without strictly exogeneous regressors. Econometrica 70:2053–2063
36. Honoré BE, Tamer E (2006) Bounds on parameters in panel dynamic discrete choice models. Econometrica 74:611–629
37. Im KS, Ahn SC, Schmidt P, Wooldridge JM (1999) Efficient estimation of panel data models with strictly exogenous explanatory variables. J Econom 93:177–201
38. Imbens GW, Wooldridge JM (2007) What's new in econometrics? Lecture Notes. National Bureau of Economic Research, Summer Institute
39. Inoue A (2008) Efficient estimation and inference in linear pseudo-panel data models. J Econom 142:449–466
40. Keane MP, Runkle DE (1992) On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. J Bus Econ Stat 10:1–9
41. Kiefer NM (1980) Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance. J Econom 14:195–202
42. Levinshohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. Rev Econ Stud 70:317–341
43. Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
44. Moon HR, Phillips PCB (2004) GMM estimation of autoregressive roots near unity with panel data. Econometrica 72:467–522
45. Mundlak Y (1961) Empirical production function free of management bias. Farm J Econ 43:44–56
46. Mundlak Y (1978) On the pooling of time series and cross section data. Econometrica 46:69–85
47. Murtazashvili I, Wooldridge JM (2007) Fixed effects instrumental variables estimation in correlated random coefficient panel data models. J Econom 142:539–552
48. Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. Econometrica 64:1263–1298

49. Pesaran MH, Takashi Y (2008) Testing slope homogeneity in large panels. J Econom 142:50–93
50. Pesaran MH, Smith RP, Im KS (1996) Dynamic linear models for heterogeneous panels. In: Mátáyas L, Sevestre P (eds) The econometrics of panel data. Kluwer, Dordrecht, pp 145–195
51. Petrin A, Train KE (2005) Tests for omitted attributes in differentiated product models. Mimeo. University of Minnesota, Department of Economics
52. Phillips PCB, Moon HR (2000) Nonstationary panel data analysis: An overview of some recent developments. Econom Rev 19:263–286
53. Wooldridge JM (1999) Distribution-free estimation of some nonlinear panel data models. J Econom 90:77–97
54. Wooldridge JM (2000) A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. Econ Lett 68:245–250
55. Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT Press, Cambridge
56. Wooldridge JM (2005) Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. Rev Econ Stat 87:385–390
57. Wooldridge JM (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J Appl Econom 20:39–54

Books and Reviews

Arellano M (2003) Panel data econometrics. Oxford University Press, Oxford

Baltagi BH (2001) Econometric analysis of panel data, vol 2e. Wiley, New York

Hsiao C (2003) Analysis of panel data, vol 2e. Cambridge University Press, Cambridge

Econophysics, Observational

BERTRAND M. ROEHNER
Institute for Theoretical and High Energy Physics,
University of Paris 7, Paris, France

Article Outline

Glossary

Cross national comparisons Comparing cross-national data for a specific phenomenon, e. g. a surge in housing prices, is the key to distinguishing between essential factors which are common to all episodes and those which are accessory and context dependent.

Economathematicians Mathematicians or theoretical physicists who develop mathematical tools, models or simulations for social phenomena but do not try to confront these models to actual observations.

Econophysics A field of physics which originated in the mid-1990s. Throughout this article, we use the term in a broad sense which includes econophysics, sociophysics and historiophysics. As a matter of fact, these fields can hardly be studied separately in the sense that economic effects depend upon social reactions (e. g. reactions of consumers to advertising campaigns); furthermore, economic investigations crucially rely on statistics which typically must combine present-day data with data from former historical episodes.

Econophysicists Physicists who study social, economic or political issues.

Endogenous mechanisms Models usually describe endogenous mechanisms. For instance a population model would describe how people get married and have children.

Exogenous factors Exogenous factors are more or less unexpected external forces which act on the system. Thus, for a population wars or epidemics may bring about sudden population changes. It is only when exogenous factors are recurrent and fairly repetitive that they can be taken into account in models.

Experiment Apart from its standard meaning in physics or biology we also use this term to designate the process of (i) defining the phenomenon that one wants to study (ii) locating and collecting the data which are best suited for the investigation (iii) analyzing the data and deriving *regularity rules* or testing a model.

Model testing Before confronting the predictions of a model to statistical evidence it is necessary to ensure that the system was not subject to unexpected exogenous shocks. The impact of exogenous factors which are not accounted for in the model must in some way be removed, that is to say the data must be corrected in a way which takes these shocks out of the picture. Usually, such corrections are very tricky to implement.

Definition of the Subject

> "No science thrives in the atmosphere of direct practical aim. We should still be without most of the conveniences of modern life if physicists had been as eager for immediate applications as most economists are and always have been." (J. Schumpeter p.6 in [11])

> "The free fall is a very trivial physical phenomenon, but it was the study of this exceedingly simple fact and its comparison with the astronomical material which brought forth mechanics. The sound procedure [in every science] is to obtain first utmost precision and mastery in a limited field, and then to proceed to another, somewhat wider one and so on." (J. von Neumann and O. Morgenstern [5])

These two quotes define fairly well the path that econophysics tries to follow. They both insist on the fact that one should begin by focusing on simple phenomena even if at first sight they have little practical implications. In what follows we will develop this point but first of all we must address a question which comes to the mind of all persons who hear about econophysics for the first time, namely:

> "Why should physicists have something to say about economic and social phenomena. Admittedly, biology can benefit from physics because of the means of observation [e. g. exploration of protein molecules by X-ray scattering] that it provides, but there are no similar needs in economics."

I have heard this question asked repeatedly by many of my colleagues. In my answer I usually emphasize that what matters is more the method of investigation than the phenomena by themselves. I stress that applying to the social sciences the experimental methodology invented by physicists and chemists would mark a great progress. However, with the benefit of insight, I realize that these answers may have appeared far fetched and unconvincing to many of my listeners. A better and more factual claim is to observe that over the past century several of the most renowned economists and sociologists were in fact econophysicists in the sense defined in the glossary. Indeed, back in the nineteenth century, the only way to get a decent mathematical training was to study astronomy, engineering, mathematics or physics. When such people entered the social sciences this lead to two kinds of approaches which we may designate as econophysics and economathematics (see Sect. "Glossary"). In the first category one may mention the astronomer Adolphe Quételet (1796–1874), Clément Juglar (1819–1905) educated as a medical doctor, Vilfredo Pareto (1848–1923) educated as an engineer, the mathematician Louis Bachelier (1870–1946), the physicist Elliott Montroll (1916–1983), the mathematician Benoît Mandelbrot (1924–). In the second category one may mention Léon Walras (1834–1910) who was educated as an engineer, the astronomer Simon Newcomb (1835–1905), the physicist Maurice Allais (1911–).

Of course, if the economic discipline had been highly successful there would be little need for an alternative approach. However, great doubts have been expressed by some of the most renowned economists about the attainments of their discipline. We have already cited Schumpeter's opinion on this matter. In addition one may mention the judgments of Vassily Leontief, Anna Schwartz, Lawrence Summers or the thesis developed in a recent book by Masanao Aoki and Hiroshi Yoshikawa.

- Leontief and Schwartz emphasized that the present organization of economic research discourages observational research. In Schwartz's words [12][1]

  > "The main disincentive to improve the handling and use of data is that the profession withholds recognition to those who devote their energies to measurement. Someone who introduces an innovation in econometrics, by contrast, will win plaudits."

- The assessment made by Summers in a paper published in 1991 is well summarized by its title: "The scientific illusion of empirical macroeconomics".
- In their book, Aoki and Yoshikawa (p. 25 in [1]) point out that the representative agent assumption which is supposed to provide a connection between micro- and macroeconomics is fundamentally flawed because it neglects both social variability and stochastic fluctuations. It may be true that in recent years a greater emphasis has been put on the issue of heterogeneity. Yet, is this the right way to takle the problem? A model is a simplification of reality anyway, so if it is not tenable to use loosely defined representative agents, an alternative solution may be to focus on sharply defined agent's attitudes. For instance, whereas without further specification home buyers may not be well defined as a useful category, the behavior of investors during the final phases of speculative price peaks may be sufficiently recurrent to make up for a well defined category.

Introduction

What are the main characteristics of econophysics? In what follows we will try to summarize some basic principles. Each of them will be illustrated by one or several studies performed by econophysicists over the past decade. Although the wording that we use is fairly personal, we believe that fundamentally these principles are shared by many econophysicists. In the course of more than a decade, econophysics has become a big tree with many branches. Obviously it is impossible to describe all of them if only because the knowledge and understanding of the present author is limited. He apologizes in advance for his limitations and for the fact that the present selection is by necessity fairly subjective.

The Primacy of Observation

Econophysics started around 1995 in sync with the creation of huge computerized databases giving minute by minute transactions on financial markets such as the New York stock market, the dollar-yen exchange rate, the forward interest rates or providing individual income data for millions of people. It may be estimated that between 1995 and 2005 about two thirds of the papers published by econophysicists aimed at deriving *regularity rules* from such databases. Let us illustrate this point by the case of income data. Since Pareto's work we know that the distribution of high incomes can be described by a power law with an exponent α comprised between 1 and 1.5. With databases comprising millions of income data one can get

[1] Leontief (p. xi in [3]) has even stronger words: "The methods used to maintain intellectual discipline in this country's most influential economics departments can occasionally remind one of those employed by the Marines to maintain discipline on Parris Island [a training camp of US Marines]."

high accuracy estimates for α and observe how α changes as the result of economic booms or stock market crashes. It turns out that α decreases during booms and increases in the wake of stock market collapses [6].

Other empirical investigations were carried out in the past decades. We list some of them below. The list is arranged by topic and by research teams.

- Stock transactions, (i) Boston University: see publications involving G. Stanley. (ii) CEA (i. e. Commissariat à l'Energie Atomique which means Institute for Atomic Research) and "Science-Finance": see publications involving J.P. Bouchaud. (iii) Nice University and UCLA: see publications involving D. Sornette. (iv) University of Warsaw: see publications involving J. Kertesz.
- Forward interest rates, Singapore University: see publications involving B. Baaquie.
- Exchange rates, Zurich: see publications involving M. Dacorogna.

To many physicists the statement that observation is supreme could seem self evident. In economics, however, such a statement represents a revolution. We already mentioned the fact that observation is a neglected topic in economics. As a matter of fact, before econophysics started it was impossible to publish a paper which would identify *regularity rules* without at the same time providing a model[2].

Investigating One Effect at a Time

In most natural phenomena different effects occur simultaneously. For instance, if one leaves a glass of cold water in the sun, the water will of course get warmer but if one looks at the mechanisms which are implied this involves many different effects: interaction of light and water, interaction of light and glass, conduction of heat, creation of convection currents between layers of water which are at different temperatures, and so on. One of the main challenges of physics was to identify these effects and to study them separately. Similarly, most social phenomena involve different effects; thus, one of the main tasks of the social sciences should be to disentangle and decompose complex phenomena into simple effects. In principle this is easier to do in physics than in the social sciences because one can change experimental conditions fairly easily. However, history shows that the main obstacle are conceptual. The previous phenomenon involves the transformation of one form of energy (light) into other forms of energy and it is well know that it took centuries for a clear understanding of these processes to emerge. In order to convince the reader that the same approach can be used in the social sciences we briefly describe a specific case.

Suicide is commonly considered as a phenomenon which is due to many factors. One of them is the strength of the marital bond. How can we isolate that factor? Of course, it is impossible to isolate it completely but one can at least make it so predominant that other factors become negligible. To achieve that objective, we consider a population in which the number of males is much larger than the number of females. Such a population will necessarily have a large proportion of bachelors and therefore will be an ideal testing ground to study the role of the marital bond. Where can we find populations with a large excess of men? Almost all populations of immigrants are characterized by an excess of males. It turns out that due to specific circumstances, this imbalance was particularly large in the population of Chinese people living in the United States. By the end of the 19th century there were about 27 Chinese men for one Chinese woman[3].

What makes the present principle important? Unless one is able to estimate the impact of each factor separately, one will never gain a *lasting* understanding. It is important to understand why. Let us for a moment return to the previous experiment. In the econometric approach one would conduct multivariate regressions of the temperature as a function of various (pre-conceived) parameters such as the volume of the liquid, the thickness of the glass and so on. Now suppose we wish to predict what happens when water is replaced by black ink. As a result of greater light absorption temperature differentials will be larger and convection currents will be stronger. The fact that many effects change at the same time will make the multivariate estimates irrelevant. Unless one has an understanding of the various individual effects it will be impossible to make any sound prediction. To sum up, any major change in business and social conditions will invalidate the previously accepted econometric models. This explains why the econometric approach fails to ensure that knowledge grows in a cumulative way.

What Guidance Can Physics Provide?

One can recall that the experimental methodology pioneered by researchers such as Tycho Brahe (1546–

[2] In what economists call "empirical econometrics" the researcher necessarily must provide a multivariate econometric model which means that even before he analyses the data he already knows the theory which rules the phenomenon. Moreover, all factors whether they have a weak or a strong impact are treated on the same footing. As we will see in the next point this has important implications.

[3] For more details about this case, see [9].

1601), Johannes Kepler (1571–1630) or Galileo (1564–1642) marked the beginning of modern physics. Two centuries later, that methodology was adapted to the exploration of the living world by people such as Claude Bernard (1813–1878), Louis Pasteur (1822–1895) and Gregor Mendel (1822–1884). In a sense it is a paradox that this method has been used successfully for the understanding of living organisms but has not yet gained broad acceptance in the social sciences for it can be argued with good reason that living organisms are more complex systems than are states or societies[4]. In short, applying the experimental methodology to the social sciences is a move which seems both natural and long overdue. Actually, serious efforts were made in this direction by social scientists such as Emile Durkheim (1858–1917) or Vilfredo Pareto (1848–1923) but this route seems to have been sidetracked in the second half of the 20th century.

Can we use the mathematical framework of physics in the investigation of social phenomena? This approach has been tried with some success by renowned econophysicists such as Belal Baaquie and coworkers (2004, 2007) and Jean-Philippe Bouchaud and coworkers [2,4]. In those cases the success must probably be attributed to the fact that the methods of theoretical physics which were used could be formulated in a purely mathematical way which did not rely on any physical concepts such as energy, momentum or temperature. As we do not yet know how these notions should be transposed to social systems, it seems impossible to apply the formalism of statistical mechanics to social phenomena[5].

Our claim that the experimental methodology of physics can be used to explore social phenomena must be substantiated by explaining how it is possible to carry out "experiments" in social phenomena. This is the purpose of the next section.

How Cross-National Observations Can Be Used to Test the Role of Different Factors

Nowadays when a solid state physicist wants to measure, say, the interaction between ultraviolet light and a crystal of germanium, the experiment involves little uncertainties. That is so because this field of physics is already well understood. On the contrary, in the case of new and not well understood phenomena there is considerable uncertainty about the specific conditions of the experimental set up. In the two years after Léon Foucault demonstrated the Foucault pendulum experiment, at least twenty physicists tried to repeat it. Some succeeded while others did not. Indeed the experimental conditions, e. g. the length of the pendulum or the nature of the suspension wire, ensuring that the Foucault effect will be observed were not well understood. It is only through various attempts with different settings that a better understanding progressively emerged. For instance it was realized that by using a pendulum of great length one would be able to reduce two undesirable effects (i) the sensitivity of the pendulum to exogenous noise[6] (ii) the Puiseux effect which generates a rotation of the oscillation plane which interferes with the Foucault effect.

Few (if any) sociological phenomena are well understood which means that social researchers are basically in the same situation as those physicists in the years 1851–1852 who tried to observe the Foucault effect[7]. As an illustration suppose we wish to know if the publication of a specific type of news has an effect on the number of suicides[8]. Such an observation depends upon many parameters: the nature of the news and the amount of attention that it receives, the time interval (days, weeks or months?) between the publication of the news and the occurrence of the suicides. In addition one does not know if there will be an increase or a decrease in the number of suicides, if men will be more or less affected than women, and so on. All these questions can in principle be answered by conducting many observations in different countries and in different periods of time. In other words, if we are sufficiently determined, patient and tenacious and if we can get access to the statistical data that are needed, we should be able to disentangle and unravel the phenomenon under consider-

[4] We will not develop this point here but it can be observed that a bacteria or a cell contains thousands of different proteins which interact in various ways. In the same line of thought one may recall that living organisms have been around for several billions years whereas societies appeared less than 100,000 years ago and states less than 10,000 years ago.

[5] It could be argued that one is free to define "social energy" in the way which one wishes. However, one should remember that the notion of energy is pivotal in physics only because it is ruled by (experimentally proved) conservation laws, such as the equivalence between heat and mechanical energy demonstrated by James Joule. Naturally, prior to defining a "social temperature", it would seem natural to define a herd- or swarm-temperature describing aggregated populations of bacteria, insects or animals. As far as we know, no operational definition of this kind has yet been proposed.

[6] Indeed, it is when the speed of the pendulum goes through zero that it is particularly sensitive to external perturbations; increasing the length of the pendulum reduces the number of oscillations in a given time interval and therefore the drift due to noise.

[7] As a more recent and even less understood case, one can mention the physicists who keep on trying to observe the cold fusion effect.

[8] This question is connected to what is known in sociology as the Werther effect; for more details see the papers written by Phillips (in particular [7]) and Chap. 3 in [9].

ation in the same way as experimenters have been able to determine how the Foucault effect can be observed.

How Vested Interests May Affect the Accessibility and Reliability of Social Data

So far we have emphasized the similarities between natural and social phenomena but there are also some stumbling blocks which are specific to the social sciences. One of them is the fact that some data may have been altered or swept under the carpet by some sort of ideological, political or social bias, pressure or interference. Needless to say, extreme care must be exercised in such cases before making use of the data.

As an illustration, suppose that an econophysicist or a sociologist wants to study episodes of military occupation of one country by another. Such episodes are of particular interest from a sociological perspective because they bring about strong interactions and can serve to probe the characteristics of a society. Moreover, because armies display many similarities no matter their country of origin, such episodes offer a set of *controlled experiments*. Naturally, in order to be meaningful the comparison must rely on trustworthy accounts for each of the episodes. Unfortunately, it turns out that in many cases only scant and fairly unreliable information is available . Consider for instance the occupation of Iceland by British and American forces during World War II. Among all occupation episodes this one was particularly massive with troops representing 50% of the population of Iceland prior to the occupation. The same proportion in a country such as Japan would have meant 30 million occupation troops that is 60 times more than the peak number of 500,000 reached at the end of 1945. Quite understandably for such a high density of troops, there were many incidents with the population of Iceland[9]; yet, is is difficult to find detailed evidence. Due to the paucity of data a superficial investigation would easily lead to the conclusion that there were in fact only few incidents. It does not require much imagination to understand why this information has not been released. The fact that in a general way all countries whatsoever are reluctant to recognize possible misconduct of their military personnel explains why the information is still classified in British and American archives. Because Iceland and the United States became close allies after 1945, one can also understand that the Icelandic National Archive is reluctant to release information about these incidents. The same observation also applies (and for the same reasons) to the occupation of Japan, 1946–1951; for more details see Roehner pp. 90–98 in [9] and [10]. Naturally, similar cases abound. Due to a variety of reasons well-meaning governments, archivists and statistical offices keep sensitive files closed to social scientists. Most often it is in fact sufficient to catalog sensitive file units in a fairly obscure way. The plain effect is that the information will not be found except perhaps by pure luck, a fairly unlikely prospect in big archives.

How Can Exogenous Factors be Taken into Account?

This question is not specific to social phenomena, it is also of importance in physics. As a matter of fact, in astronomy it provides a powerful method for observing objects that cannot be observed directly. Thus, we know the existence of exoplanets only from the perturbing effect which they have on the position of the star around which they move. However, for social phenomena the problem of exogenous factors is much more serious because (i) they may not be known to observers (ii) even once they are identified it is very difficult to correct the data in a reliable way. One of the main pitfalls in the modeling of socio-economic phenomena is to explain them through endogenous mechanisms while they are in fact due to exogenous factors. The following examples make clear that this difficulty exists for many phenomena, whether they belong to the financial, economic or social sphere.

- In their paper of 2005 about consensus formation and shifts in opinion Michard and Bouchaud confront their theory to two classes of social phenomena: (i) the diffusion of cell phones (ii) the diffusion of birth rate patterns. In the first case it is clear that advertising campaigns may have played an important role. Of course, one could argue that these campaigns were part of the endogenous diffusion process. However, this argument does not hold for big telecom companies (e. g. Vodafone) which operate in many countries. In such cases the decision about the magnitude of the advertising campaigns are taken by the board of the company which means that such campaigns can hardly be considered as endogenous effects. Similarly, birth rates depend upon exogenous factors. For instance the length of time spent in higher education has an effect on the average age of marriage and the later has an effect on birth rates.
- On 21 July 2004 the share price of Converium, a Swiss reinsurance company listed on the New York Stock

[9] According to a report that Prime Minister Hermann Jonasson sent to the American Headquarters, there were 136 incidents between troops and Icelanders during the period between July 1941 (arrival of the American troops) and April 1942 (Hunt 1966) in Reykjavik alone. Unfortunately, no copy of this report seems to be available at the National Archives of Iceland.

Exchange dropped 50%. Was this fall the result of an avalanche effect due to a movement of panic among investors? In fact, the most likely explanation is that it was the consequence of a decision taken by the board of Fidelity International, a major investment fund and one of the main shareholders of Converium. Indeed in a statement issued by Converium on August 3, 2004 it was announced that Fidelity had reduced its holdings from 9.87% to 3.81%. In other words, it would be completely irrelevant to explain such a fall through a herd effect model or through any other endogenous mechanism (more details can be found in [8]). Similar conclusions apply to corporate stock buybacks, as well as to mergers, acquisitions, buyouts and takeovers; in all these cases decisions taken by a few persons (the average board of directors has nine members) may trigger substantial changes in share prices. How should such effects be taken into account by stock market models?

- At the end of 2004 and in the first months of 2005 British housing prices began to decline after having risen rapidly during several years. Yet after May 2005, they suddenly began to pick up again at an annual rate of about 10%. This resurgence was particularly intriguing because at the same time US housing prices began to decline. To what factor should this unexpected rise be attributed? Most certainly this was the market response to a plan introduced by the Chancellor of the Exchequer Gordon Brown in late May (The Economist May 28, 2005). Under this plan which aimed at propping up house prices new buyers would benefit from a zero-interest loan for 12% of the price. In addition, the government would cover all losses incurred by banks as a result of possible bankruptcies of borrowers (at least so long as prices did not fall by more than 12%). It appears that the plan indeed propped up the market. Consequently, in order to confront the predictions of any model (e. g. see Richmond's paper which was published in 2007) with observation the impact of this plan effect must first be taken out of the picture.
- The same difficulty is also encountered in socio-political phenomena. Here is an illustration. On 5 October 2000, in protest against the publication of the results of the presidential election there was a huge mass demonstration in Belgrade which involved thousands of people from the provinces who were transported to the federal capital by hundreds of buses. It clearly showed that president Milosevic was no longer in control of the police and army and lead to his retirement from the political scene. Thus, what NATO air strikes (24 March-11 June 1999[10]) had not been able to achieve was accomplished by one night of street demonstrations. What was the part of exogenous factors in this event? Although in many similar cases it is very difficult to know what really happened, in this specific case a partial understanding is provided by a long article published in the New York Times[11]. In this article we learn that several American organizations belonging to the intelligence network supported, financed and trained Serbian opposition groups. For instance the article mentions the Albert Einstein Foundation, the International Republican Institute, the National Endowment for Democracy, the US Agency for International Development. Although the amount of the total financial support is not known, the New York Times article says that it exceeded $ 28 million. The plan comprised two facets: the organization of demonstrations on the one hand and the infiltration of the army and police on the other hand in order to undermine their loyalty and convince them to remain passive during the demonstrations. According to the article this second facet remains classified. With an exogenous interference of such a magnitude, it would clearly be meaningless to describe this upheaval as a purely endogenous process. Moreover, the fact that we have only partial knowledge about the exogenous forces makes it very difficult (if not altogether impossible) to come up with a satisfactory description. It should also be noted that the influence of these groups did not disappear overnight after October 4, which means that the subsequent history of Serbia must also take them into account at least to some extent.

Future Directions

In this article we have described the challenges and obstacles to which one is confronted in trying to understand socio-economic phenomena. In parallel we have shown that the econophysics approach has many assets. One of them which has not yet been mentioned is the fact that econophysicists are not subject to the rigid barriers which ex-

[10] It can be noted that similarly to what would happen in 2003 for the invasion of Iraq, these air strikes were carried out without the authorization of the United Nations Security Council.

[11] New York Times, Sunday 26 November 2000, Magazine Section, p. 43, 7705 words; the article by Roger Cohen is entitled: "Who really brought down Milosevic". What makes this account particularly convincing is the fact that it was preceded by another article entitled:"US anti-Milosevic plan faces major test at polls" which appeared on September 23, 2000 (p. 6, 1150 words); this article described the way Milosevic would be removed from power two weeks *before* the events. The article makes clear that the course of events would be the same no matter what the results of the election would be.

ist between various fields and subfields of the human sciences. Thus, if it turns out that in order to explain an economic phenomena one needs to understand a social effect, econophysicists would have no problem in shifting from one field to another. There is another historical chance that we have not mentioned so far, namely the development of the Internet. In the past decade 1997–2007 the amount of information to which one has access has increased tremendously. Electronic catalogs of major libraries or of national archives, indexes of newspaper, search engines on the Internet, searchable databases of books, all these innovations contributed to give the researcher easy access to information sources that have never been available before. In particular it has become fairly easy to find cross-national data. Thus, social scientists and econophysicists are in a better position than ever for carrying out the kind of comparative studies that we called for in this article.

Bibliography

Primary Literature

1. Aoki M, Yoshikawa H (2007) Reconstructing macroeconomics. Cambridge University Press, Cambridge
2. Bouchaud J-P, Potters M (2003) Theory of financial risk and derivative pricing. Cambridge University Press, Cambridge
3. Leontief W (1983) Foreword. In: Eichner AS (ed) Why economics is not yet a science. M.E. Sharpe, Armonk(New York)
4. Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. Eur. Phys. J. B 47:151–159
5. Neumann J von, Morgenstern O (1953) Theory of games and economic behavior. Princeton University Press, Princeton
6. Nirei M, Souma W (2007) Two factor model of income distribution dynamics. Review of Income and Wealth 53(3):440–459
7. Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am. Sociol. Rev. 39:340–354
8. Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) Proceedings of the 3rd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 262–271
9. Roehner BM (2007) Driving forces in physical, biological and socio-economic phenomena. Cambridge University Press, Cambridge
10. Roehner BM (2008) Relations between Allied forces and the population of Japan, Working Report UPMC, Paris
11. Schumpeter J (1933) The common sense of econometrics. Econometrica 1:5–12
12. Schwartz AJ (1995) An interview with Anna J. Schwartz. Newsl. Cliometric Soc. 10(2):3–7

Books and Reviews

Two observations are in order about this reference section:

Many of these references are not mentioned in the text; the objective is to give readers a starting point for further readings on various aspects of econophysics.

There is a fairly complete list of publications of the present author; it is given for the purpose of illustrating through one specific case the "trajectory" of an econophysicist in the course of time (1995–2007).

Amaral LAN, Buldyrev SV, Havlin S, Leschhorn H, Maass P, Salinger A, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: I. Empirical results for company growth. J Phys. I Fr. 7:621–633

Amaral LAN, Buldyrev SV, Havlin S, Salinger MA, Stanley HE (1998) Power law scaling for a system of interacting units with complex internal structure. Phys. Rev. Lett. 80(7):1385–1388

Aoki M, Yoshikawa H (2007) Reconstructing macroeconomics. Cambridge University Press, Cambridge

Baaquie BE (2004) Quantum finance. Cambridge University Press, Cambridge

Baaquie BE (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. I. Theory Phys. Rev. E 75, 016703

Baaquie BE, Liang C (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. II. Empirical Phys. Rev. E 75, 016704

Baaquie BE, Srikant M (2004) Comparison of field theory models of interest rates with market data. Phys. Rev. E 69, 036129

Borghesi C, Bouchaud J-P (2007) On songs and men. Quality and Quantity 41(4):557–568

Bouchaud J-P, Marsili M, Roehner BM, Slanina F (eds) (2001) Application of physics in economic modelling. Proceedings of the NATO Advanced Research Workshop held in Prague, Czech Republic, 8–10 February 2001. Physica A 299(1–2):1–355

Bouchaud J-P, Potters M (1997) Théorie des risques financiers. Aléa, Saclay

Bouchaud J-P, Potters M (2003) Theory of financial risk and derivative pricing. Cambridge University Press, Cambridge

Buldyrev SV, Amaral LAN, Havlin S, Leschhorn H, Maass P, Salinger MA, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: II. Modeling of company growth. J Phys. I Fr. 7:635–650

Chakraborti A, Chakrabarti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. Eur. Phys. J. B 17:167–170

de Oliveira SM,de Oliveira PMC, Stauffer D (1999) Evolution, money, war and computer. Teubner, Leipzig

Deschâtres F, Sornette D (2005) The dynamics of book sales: endogenous versus exogenous shocks in complex networks. Phys. Rev. E 72, 016112

Dragulescu A, Yakovenko VM (2000) Statistical mechanics of money. Eur. Phys. J. B 17(4):723–729

Farmer JD (1999) Physicists attempt to scale the ivory towers of finance. Comput. Sci. Eng. Nov-Dec 1999, 26–39

Farmer JD, Lillo F (2004) On the origin of power law tails in price fuctuations. Quant. Finance 4(1):7–11

Feigenbaum JA, Freund PGO (1998) Discrete scale invariance and the second Black Monday. Mod. Phys. Lett. B 12(2–3):57–60

Fu Y-Q, Zhang H, Cao Z, Zheng B, Hu G (2005) Removal of pinned spiral by generating target waves with a localized stimulus. Phys. Rev. E 72, 046206

Galam S (2006) Opinion dynamics, minority spreading and heterogenous beliefs. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) Econophysics and Sociophysics. Wiley-VCH, Weinheim

Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. Nature 381:767–770

Guillaume DM, Dacorogna MM, Davé R, Müller UA, Olsen RB, Pictet OV (1997) From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets. Finance Stoch. 1:95–129

Hunt JJ (1966) The United States occupation of Iceland, 1941–1946. Thesis. Georgetown University, Washington DC

Johansen A, Sornette D (1999) Financial anti-bubbles: Log-periodicity in gold and Nikkei collapses. Int. J. Mod Phys C 10(4):563–575

Johansen A, Sornette D (2001) Bubbles and anti-bubbles in Latin-American, Asian and Western stock markets: An empirical study. Int. J. Theor. Appl. Finance 4(6):853–920

Juglar C (1862): Des crises commerciales et de leur retour périodique en France, en Angleterre et aux Etats-Unis. English translation (1893, 1966) A brief history of panics and their periodical occurrence in the United States. A.M. Kelley, New York

Lai KK, Leung FKN, Tao B, Wang S (2000) Practices of preventive maintenance and replacement for engines: a case study. Eur J Oper Res 124:2

Leontief W (1983) Foreword. In: Eichner AS (ed) Why economics is not yet a science. M.E. Sharpe, Armonk New York

Li M, Wu J, Wang D, Zhou T, Di Z, Fan Y (2006) Evolving model of weighted networks inspired by scientific collaboration networks. Physica A 375(1):355–364

Lillo F, Mike S, Farmer JD (2005) Theory for Long Memory in supply and semand. Physical Review E 7106 (6 pt 2) 287–297

Lux T (1996) The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. Appl Financial Econ 6:463–475

Mandelbrot B (1997) Les fractales et la Bourse. Pour Sci 242:16–17

Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B 11:193–197

Mantegna RN, Stanley HE (1995) Scaling behavior in the dynamics of an economic index. Nature 376:46–49

Mantegna RN, Stanley HE (1999) Introduction to econophysics. Cambridge University Press, Cambridge

McCauley JL (2004) Dynamics of markets. Cambridge University Press, Cambridge

Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. Eur Phys J B 47:151–159

Mimkes J (2006) A thermodynamic formulation of social science. In: Chakrabarti BK, Chaterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Weinheim, pp 279–310

Müller UA, Dacorogna MM, Davé R, Olsen RB, Pictet OV, Weizsäcker J von (1997) Volatilities of different time resolutions. Analysing the dynamics of market components. J Empir Finance 4(2–3):213–240

Müller UA, Dacorogna MM, Olsen RB, Pictet OV, Schwarz M (1990) Statistical study of foreign exchange rates, empirical evidence of a price scaling law, and intraday analysis. J Bank Finance 14:1189–1208

Neumann J von, Morgenstern O (1953) Theory of games and economic behavior. Princeton University Press, Princeton

Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am Sociol Rev 39:340–354

Plerou V, Amaral LAN, Gopikrishnan P (1999) Similarities between the growth dynamics of university research and of competitive economic activities. Nature 400(6743):433–437

Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Stanley H (1999) Universal and nonuniveral properties of cross-correlation in financial time series. Phys Rev Lett 83(7):1471–1474

Richmond P (2007) A roof over your head; house price peaks in the UK and Ireland. Physica A 375(1,15):281–287

Roehner BM (1995) Theory of markets. Trade and space-time patterns of price fluctuations: a study in analytical economics. Springer, Berlin

Roehner BM (1997) Jesuits and the state. A comparative study of their expulsions (1500–1990). Religion 27:165–182

Roehner BM (1997) The comparative way in economics: a reappraisal. Econom Appl 50(4):7–32

Roehner BM (1999) Spatial analysis of real estate price bubbles: Paris 1984–1993. Reg Sci Urban Econ 29:73–88

Roehner BM (1999) The space-time pattern of price waves. Eur Phys J B 8:151–159

Roehner BM (2000) Determining bottom price-levels after a speculative peak. Eur Phys J B 17:341–345

Roehner BM (2000) Identifying the bottom line after a stock market crash. Int J Mod Phys C 11(1):91–100

Roehner BM (2000) Speculative trading: the price multiplier effect. Eur Phys J B 14:395–399

Roehner BM (2000) The correlation length of commodity markets: 1. Empirical evidence. Eur Phys J B 13:175–187

Roehner BM (2000) The correlation length of commodity markets: 2. Theoretical framework. Eur Phys J B 13:189–200

Roehner BM (2001) Hidden collective factors in speculative trading: a study in analytical economics. Springer, Berlin

Roehner BM (2001) To sell or not to sell? Behavior of shareholders during price collapses. Int J Mod Phys C 12(1):43–53

Roehner BM (2001) Two classes of speculative peaks. Physica A 299:71–83

Roehner BM (2002) Patterns of speculation: a study in observational econophysics. Cambridge University Press, Cambridge

Roehner BM (2002) Patterns and repertoire. Harvard University Press, Cambridge Massachussets

Roehner BM (2002) Separatism and integration. Rowman and Littlefield, Lanham Maryland

Roehner BM (2004) Patterns of speculation in real estate and stocks. In: Takayasu H (ed) Proceedings of the 2nd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 103–116

Roehner BM (2005) A bridge between liquids and socio-economic systems: the key-role of interaction strengths. Physica A 348:659–682

Roehner BM (2005) Cohésion sociale. Odile Jacob, Paris

Roehner BM (2005) Stock markets are not what we think they are: the key roles of cross-ownership and corporate treasury stock. Physica A 347:613–626

Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) Proceedings of the 3rd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 262–271

Roehner BM (2006) Real estate price peaks: a comparative perspective. Evol Institutional Econ Rev 2(2):167–182

Roehner BM (2007) Driving forces in physical, biological and socio-economic phenomena. Cambridge University Press, Cambridge

Roehner BM (2008) Econophysics: Challenges and promises. Evolutionary abd Institutional Economics Review 4(2):251–266

Roehner BM, Jego C (2006) White flight or flight from poverty? J Econ Intearct Coord 1:75–87

Roehner BM, Maslov S (2003) Does the price multiplier effect also hold for stocks? Int J Mod Phys C 14(10):1439–1451

Roehner BM, Maslov S (2003) The conundrum of stock versus bond prices. Physica A 335:164–182 (2004)

Roehner BM, Rahilly LJ (2002) Separatism and integration: a study in analytical history. Rowman and Littlefield, Lanham Maryland

Roehner BM, Shiue C (2001) Comparing the correlation length of grain markets in China and France. Int J Mod Phys C 11(7):1383–1410

Roehner BM, Sornette D (1998) The sharp peak – flat trough pattern and critical speculation. Eur Phys J B 4:387–399

Roehner BM, Sornette D (1999) Analysis of the phenomenon of speculative trading in one of its basic manifestations: postage stamp bubbles. Int J Mod Phys C 10(6):1099–1116

Roehner BM, Sornette D (2000) Thermometers of speculative frenzy. Eur Phys J B 16:729–739

Roehner BM, Sornette D, Andersen J (2004) Response functions to critical shocks in social sciences: an empirical and numerical study. Int J Mod Phys C 15(6):809–834

Roehner BM, Syme T (2002) Pattern and repertoire in history: an introduction to analytical history. Harvard University Press, Cambridge Massachusetts

Schumpeter J (1933) The common sense of econometrics. Econometrica 1:5–12

Schwartz AJ (1995) An interview with Anna J Schwartz. Newsl Cliometric Soc 10(2):3–7

Sornette D (2003) Why stock markets crash. Critical events in complex financial systems. Princeton University Press, Princeton

Stauffer D, Sornette D (1999) Self-organized percolation model for stock market fluctuations. Physica A 271(3–4):496–506

Summers LH (1991) The scientific illusion in empirical macroeconomics. Scand J Econ 93(2):129–148

Takayasu H (ed) (2004) The application of econophysics. In: Proceedings of the 2nd Nikkei Econophysics Symposium. Springer, Tokyo

Takayasu H (ed) (2006) Practical fruits of econophysics. In: Proceedings of the 3rd Nikkei Econophysics Symposium. Springer, Tokyo

Turchin P (2003) Historical dynamics. Why states rise and fall. Princeton University Press, Princeton

Wyatt M, Bouchaud J-P (2003) Self referential behaviour, overreaction and conventions in financial markets. Cond-mat/03033584

Zhou W-X, Sornette D (2003) 2000–2003 real estate bubble in the UK and not in the USA. Physica A 329(1–2):249–263

Zhou W-X, Sornette D (2003) Evidence of a worldwide stock market log-periodic anti-bubble since mid-2000. Physica A 330:543–583

Zhou W-X, Sornette D (2004) Antibubble and Prediction of China's stock market and Real-Estate. Physica A 337(1–2):243–268

Econophysics, Statistical Mechanics Approach to

VICTOR M. YAKOVENKO
Department of Physics, University of Maryland,
College Park, USA

"Money, it's a gas." Pink Floyd

Article Outline

Glossary

Probability density $P(x)$ is defined so that the probability of finding a random variable x in the interval from x to $x + \mathrm{d}x$ is equal to $P(x)\,\mathrm{d}x$.

Cumulative probability $C(x)$ is defined as the integral $C(x) = \int_x^\infty P(x)\mathrm{d}x$. It gives the probability that the random variable exceeds a given value x.

The Boltzmann–Gibbs distribution gives the probability of finding a physical system in a state with the energy ε. Its probability density is given by the exponential function (1).

The Gamma distribution has the probability density given by a product of an exponential function and a power-law function, as in (9).

The Pareto distribution has the probability density $P(x) \propto 1/x^{1+\alpha}$ and the cumulative probability $C(x) \propto 1/x^{\alpha}$ given by a power law. These expressions apply only for high enough values of x and do not apply for $x \to 0$.

The Lorenz curve was introduced by American economist Max Lorenz to describe income and wealth inequality. It is defined in terms of two coordinates $x(r)$ and $y(r)$ given by (19). The horizontal coordinate $x(r)$ is the fraction of the population with income below r, and the vertical coordinate $y(r)$ is the fraction of income this population accounts for. As r changes from 0 to ∞, x and y change from 0 to 1, parametrically defining a curve in the (x, y)-plane.

The Gini coefficient G was introduced by the Italian statistician Corrado Gini as a measure of inequality in a society. It is defined as the area between the Lorenz curve and the straight diagonal line, divided by the area of the triangle beneath the diagonal line. For perfect equality (everybody has the same income or wealth) $G = 0$, and for total inequality (one person has all income or wealth, and the rest have nothing) $G = 1$.

The Fokker–Planck equation is the partial differential equation (22) that describes evolution in time t of the probability density $P(r, t)$ of a random variable r experiencing small random changes Δr during short time intervals Δt. It is also known in mathematical literature as the Kolmogorov forward equation. The diffusion equation is an example of the Fokker–Planck equation.

Definition of the Subject

Econophysics is an interdisciplinary research field applying methods of statistical physics to problems in economics and finance. The term "econophysics" was first introduced by the prominent theoretical physicist Eugene Stanley in 1995 at the conference *Dynamics of Complex Systems*, which was held in Calcutta (now known as Kolkata) as a satellite meeting to the STATPHYS-19 conference in China [1,2]. The term appeared in print for the first time in the paper by Stanley et al. [3] in the proceedings of the Calcutta conference. The paper presented a manifesto of the new field, arguing that "behavior of large numbers of humans (as measured, e. g., by economic indices) might conform to analogs of the scaling laws that have proved useful in describing systems composed of large numbers of inanimate objects" [3]. Soon the first econophysics conferences were organized: *International Workshop on Econophysics*, Budapest, 1997 and *International Workshop on Econophysics and Statistical Finance*, Palermo, 1998 [2], and the book *An Introduction to Econophysics* [4] was published.

The term "econophysics" was introduced by analogy with similar terms, such as "astrophysics", "geophysics", and"biophysics", which describe applications of physics to different fields. Particularly important is the parallel with biophysics, which studies living creatures, which still obey the laws of physics. It should be emphasized that econophysics does not literally apply the laws of physics, such as Newton's laws or quantum mechanics, to humans, but rather uses mathematical methods developed in statistical physics to study statistical properties of complex economic systems consisting of a large number of humans. So, it may be considered as a branch of applied theory of probabilities. However, statistical physics is distinctly different from mathematical statistics in its focus, methods, and results.

Originating from physics as a quantitative science, econophysics emphasizes quantitative analysis of large amounts of economic and financial data, which became increasingly available with the massive introduction of computers and the Internet. Econophysics distances itself from the verbose, narrative, and ideological style of political economy and is closer to **econometrics** in its focus. Studying mathematical models of a large number of interacting economic agents, econophysics has much common ground with the **agent-based modeling and simulation**. Correspondingly, it distances itself from the representative-agent approach of traditional economics, which, by definition, ignores statistical and heterogeneous aspects of the economy.

Two major directions in econophysics are applications to finance and economics. Observational aspects are covered in the article ► Econophysics, Observational. The present article, ► Econophysics, Statistical Mechanics Approach to, concentrates primarily on statistical distributions of money, wealth, and income among interacting economic agents.

Another direction related to econophysics has been advocated by the theoretical physicist Serge Galam since the early 1980s under the name“**sociophysics**” [5], with the first appearance of the term in print in [6]. It echoes the term *physique sociale* proposed in the nineteenth century by Auguste Comte, the founder of sociology. Unlike econophysics, the term “sociophysics” did not catch on when first introduced, but it is coming back with the popularity of econophysics and active promotion by some physicists [7,8,9]. While the principles of both fields have a lot in common, econophysics focuses on the narrower subject of economic behavior of humans, where more quantitative data are available, whereas sociophysics studies a broader range of social issues. The boundary between econophysics and sociophysics is not sharp, and the two fields enjoy a good rapport [10]. A more detailed description of the historical development in presented in Sect. “Historical Introduction”.

Historical Introduction

Statistical mechanics was developed in the second half of the nineteenth century by James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs. These physicists believed in the existence of atoms and developed mathematical methods for describing their statistical properties, such as the probability distribution of velocities of molecules in a gas (the Maxwell–Boltzmann distribution) and the general probability distribution of states with different energies (the Boltzmann–Gibbs distribution). There are interesting connections between the development of statistical physics and statistics of social phenomena, which were recently brought up by the science journalist Philip Ball [11,12].

Collection and study of “social numbers”, such as the rates of death, birth, and marriage, has been growing progressively since the seventeenth century (see Chap. 3 in [12]). The term “statistics” was introduced in the eighteenth century to denote these studies dealing with the civil “states”, and its practitioners were called“statists”. Popularization of social statistics in the nineteenth century is particularly accredited to the Belgian astronomer Adolphe Quetelet. Before the 1850s, statistics was considered an empirical arm of political economy, but then it started to transform into a general method of quantitative analysis suitable for all disciplines. It stimulated physicists to develop statistical mechanics in the second half of the nineteenth century.

Rudolf Clausius started development of the kinetic theory of gases, but it was James Clerk Maxwell who made a decisive step of deriving the probability distribution of velocities of molecules in a gas. Historical studies show (see Chap. 3 in [12]) that, in developing statistical mechanics, Maxwell was strongly influenced and encouraged by the widespread popularity of social statistics at the time. This approach was further developed by Ludwig Boltzmann, who was very explicit about its origins (see p. 69 in [12]):

> “The molecules are like individuals, … and the properties of gases only remain unaltered, because the number of these molecules, which on the average have a given state, is constant.”

In his book *Populäre Schriften* from 1905 [13], Boltzmann praises Josiah Willard Gibbs for systematic development of statistical mechanics. Then, Boltzmann says (cited from [14]):

> “This opens a broad perspective if we do not only think of mechanical objects. Let's consider to apply this method to the statistics of living beings, society, sociology and so forth.”

(The author is grateful to Michael E. Fisher for bringing this quote to his attention.)

It is worth noting that many now-famous economists were originally educated in physics and engineering. Vilfredo Pareto earned a degree in mathematical sciences and a doctorate in engineering. Working as a civil engineer, he collected statistics demonstrating that distributions of income and wealth in a society follow a power law [15].

He later became a professor of economics at Lausanne, where he replaced Léon Walras, also an engineer by education. The influential American economist Irving Fisher was a student of Gibbs. However, most of the mathematical apparatus transferred to economics from physics was that of Newtonian mechanics and classical thermodynamics [16]. It culminated in the neoclassical concept of mechanistic equilibrium where the "forces" of supply and demand balance each other. The more general concept of statistical equilibrium largely eluded mainstream economics.

With time, both physics and economics became more formal and rigid in their specializations, and the social origin of statistical physics was forgotten. The situation is well summarized by Philip Ball (see p. 69 in [12]):

> "Today physicists regard the application of statistical mechanics to social phenomena as a new and risky venture. Few, it seems, recall how the process originated the other way around, in the days when physical science and social science were the twin siblings of a mechanistic philosophy and when it was not in the least disreputable to invoke the habits of people to explain the habits of inanimate particles".

Some physicists and economists attempted to connect the two disciplines during the twentieth century. The theoretical physicist Ettore Majorana argued in favor of applying the laws of statistical physics to social phenomena in a paper published after his mysterious disappearance [17]. The statistical physicist Elliott Montroll coauthored the book *Introduction to Quantitative Aspects of Social Phenomena* [18]. Several economists applied statistical physics to economic problems [19,20,21,22]. An early attempt to bring together the leading theoretical physicists and economists at the Santa Fe Institute was not entirely successful [23]. However, by the late 1990s, the attempts to apply statistical physics to social phenomena finally coalesced into the robust movements of econophysics and sociophysics, as described in Sect. "Definition of the Subject".

The current standing of econophysics within the physics and economics communities is mixed. Although an entry on econophysics has appeared in the *New Palgrave Dictionary of Economics* [24], it is fair to say that econophysics is not accepted yet by mainstream economics. Nevertheless, a number of open-minded, nontraditional economists have joined this movement, and the number is growing. Under these circumstances, econophysicists have most of their papers published in physics journals. The journal *Physica A: Statistical Mechanics and Its Applications* emerged as the leader in econophysics publications and has even attracted submissions from some bona fide economists. The mainstream physics community is generally sympathetic to econophysics, although it is not uncommon for econophysics papers to be rejected by *Physical Review Letters* on the grounds that "it is not physics". There are regular conferences on econophysics, such as *Applications of Physics in Financial Analysis* (sponsored by the European Physical Society), *Nikkei Econophysics Symposium*, and *Econophysics Colloquium*. Econophysics sessions are included in the annual meetings of physical societies and statistical physics conferences. The overlap with economics is the strongest in the field of agent-based simulation. Not surprisingly, the conference series WEHIA/ESHIA, which deals with heterogeneous interacting agents, regularly includes sessions on econophysics.

Statistical Mechanics of Money Distribution

When modern econophysics started in the middle of the 1990s, its attention was primarily focused on analysis of financial markets. However, three influential papers [25,26,27], dealing with the subject of money and wealth distributions, were published in 2000. They started a new direction that is closer to economics than finance and created an expanding wave of follow-up publications. We start reviewing this subject with [25], whose results are the most closely related to the traditional statistical mechanics in physics.

The Boltzmann–Gibbs Distribution of Energy

The fundamental law of equilibrium statistical mechanics is the Boltzmann–Gibbs distribution. It states that the probability $P(\varepsilon)$ of finding a physical system or subsystem in a state with the energy ε is given by the exponential function

$$P(\varepsilon) = c e^{\frac{-\varepsilon}{T}} , \tag{1}$$

where T is the temperature, and c is a normalizing constant [28]. Here we set the Boltzmann constant k_B to unity by choosing the energy units for measuring the physical temperature T. Then, the expectation value of any physical variable x can be obtained as

$$\langle x \rangle = \frac{\sum_k x_k \mathrm{e}^{\frac{-\varepsilon_k}{T}}}{\sum_k \mathrm{e}^{\frac{-\varepsilon_k}{T}}} , \tag{2}$$

where the sum is taken over all states of the system. Temperature is equal to the average energy per particle: $T \sim \langle \varepsilon \rangle$, up to a numerical coefficient of the order of 1.

Equation (1) can be derived in different ways [28]. All derivations involve the two main ingredients: statistical character of the system and conservation of energy ε. One of the shortest derivations can be summarized as follows. Let us divide the system into two (generally unequal) parts. Then, the total energy is the sum of the parts, $\varepsilon = \varepsilon_1 + \varepsilon_2$, whereas the probability is the product of probabilities, $P(\varepsilon) = P(\varepsilon_1)P(\varepsilon_2)$. The only solution of these two equations is the exponential function (1).

A more sophisticated derivation, proposed by Boltzmann himself, uses the concept of entropy. Let us consider N particles with total energy E. Let us divide the energy axis into small intervals (bins) of width $\Delta\varepsilon$ and count the number of particles N_k having energies from ε_k to $\varepsilon_k + \Delta\varepsilon$. The ratio $N_k/N = P_k$ gives the probability for a particle having the energy ε_k. Let us now calculate the multiplicity W, which is the number of permutations of the particles between different energy bins such that the occupation numbers of the bins do not change. This quantity is given by the combinatorial formula in terms of the factorials

$$W = \frac{N!}{N_1!N_2!N_3!\ldots} . \tag{3}$$

The logarithm of multiplicity of called the entropy $S = \ln W$. In the limit of large numbers, the entropy per particle can be written in the following form using the Stirling approximation for the factorials:

$$\frac{S}{N} = -\sum_k \frac{N_k}{N} \ln\left(\frac{N_k}{N}\right) = -\sum_k P_k \ln P_k . \tag{4}$$

Now we would like to find what distribution of particles between different energy states has the highest entropy, i. e. the highest multiplicity, provided that the total energy of the system, $E = \sum_k N_k\varepsilon_k$, has a fixed value. Solution of this problem can be easily obtained using the method of Lagrange multipliers [28], and the answer gives the exponential distribution (1).

The same result can be derived from the **ergodic theory**, which says that the many-body system occupies all possible states of a given total energy with equal probabilities. Then it is straightforward to show [29,30] that the probability distribution of the energy of an individual particle is given by (1).

Conservation of Money

The derivations outlined in Sect. "The Boltzmann–Gibbs Distribution of Energy" are very general and use only the statistical character of the system and the conservation of energy. So, one may expect that the exponential Boltzmann–Gibbs distribution (1) may apply to other statistical systems with a conserved quantity.

The economy is a big statistical system with millions of participating agents, so it is a promising target for applications of statistical mechanics. Is there a conserved quantity in economy? Drăgulescu and Yakovenko [25] argue that such a conserved quantity is money m. Indeed, the ordinary economic agents can only receive money from and give money to other agents. They are not permitted to "manufacture" money, e. g., to print dollar bills. When one agent i pays money Δm to another agent j for some goods or services, the money balances of the agents change as follows:

$$\begin{aligned} m_i &\rightarrow m'_i = m_i - \Delta m , \\ m_j &\rightarrow m'_j = m_j + \Delta m . \end{aligned} \tag{5}$$

The total amount of money of the two agents before and after the transaction remains the same,

$$m_i + m_j = m'_i + m'_j , \tag{6}$$

i. e., there is a local conservation law for money. The rule (5) for the transfer of money is analogous to the transfer of energy from one molecule to another in molecular collisions in a gas, and (6) is analogous to conservation of energy in such collisions.

Addressing some misunderstandings developed in economic literature [31,32,33,34], we should emphasize that, in the model of [25], the transfer of money from one agent to another happens voluntarily, as a payment for goods and services in a market economy. However, the model only keeps track of money flow, and does not keep track of what kinds of goods and service are delivered. One reason for this is that many goods, e. g., food and other supplies, and most services, e. g., getting a haircut or going to a movie, are not tangible and disappear after consumption. Because they are not conserved and also because they are measured in different physical units, it is not very practical to keep track of them. In contrast, money is measured in the same unit (within a given country with a single currency) and is conserved in transactions, so it is straightforward to keep track of money flow.

Unlike ordinary economic agents, a central bank or a central government can inject money into the economy. This process is analogous to an influx of energy into a system from external sources, e. g., the Earth receives energy from the Sun. Dealing with these situations, physicists start with an idealization of a closed system in thermal equilibrium and then generalize to an open system subject to an energy flux. As long as the rate of money influx from central sources is slow compared with relaxation processes in the economy and does not cause hyperinflation, the system is in quasi-stationary statistical equilibrium with slowly changing parameters. This situation is analogous to heating a kettle on a gas stove slowly, where the kettle has a well-defined, but slowly increasing temperature at any moment of time.

Another potential problem with conservation of money is debt. This issue is discussed in more detail in Sect. "Models with Debt". As a starting point, Drăgulescu and Yakovenko [25] first considered simple models, where debt is not permitted. This means that money balances of agents cannot go below zero: $m_i \geq 0$ for all i. Transaction (5) takes place only when an agent has enough money to pay the price, $m_i \geq \Delta m$, otherwise the transaction does not take place. If an agent spends all the money, the balance drops to zero $m_i = 0$, so the agent cannot buy any goods from other agents. However, this agent can still produce goods or services, sell them to other agents, and receive money for them. In real life, cash balance dropping to zero is not at all unusual for people who live from paycheck to paycheck.

The conservation law is the key feature for the successful functioning of money. If the agents were permitted to "manufacture" money, they would be printing money and buying all goods for nothing, which would be a disaster. The physical medium of money is not essential, as long as the conservation law is enforced. Money may be in the form of paper cash, but today it is more often represented by digits in computerized bank accounts. The conservation law is the fundamental principle of accounting, whether in the single-entry or in the double-entry form. More discussion of banks and debt is given in Sect. "Models with Debt".

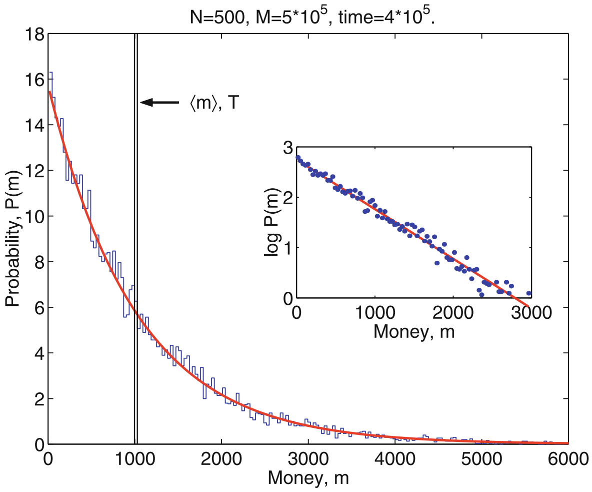
The Boltzmann–Gibbs Distribution of Money

Having recognized the principle of money conservation, Drăgulescu and Yakovenko [25] argued that the stationary distribution of money should be given by the exponential Boltzmann–Gibbs function analogous to (1):

$$P(m) = c e^{\frac{-m}{T_m}} . \tag{7}$$

Here c is a normalizing constant, and T_m is the "money temperature", which is equal to the average amount of money per agent: $T = \langle m \rangle = M/N$, where M is the total money, and N is the number of agents.

To verify this conjecture, Drăgulescu and Yakovenko [25] performed agent-based computer simulations of money transfers between agents. Initially all agents were given the same amount of money, say, \$ 1000. Then, a pair of agents (i, j) were randomly selected, the amount Δm was transferred from one agent to another, and the process was repeated many times. Time evolution of the probability distribution of money $P(m)$ can be seen in computer animation videos at the Web pages [35,36]. After a transitory period, money distribution converges to the stationary form shown in Fig. 1. As expected, the distribution is very well fitted by the exponential function (7).



Econophysics, Statistical Mechanics Approach to, Figure 1
Stationary probability distribution of money *P*(*m*) obtained in agent-based computer simulations. *Solid curves*: fits to the Boltzmann–Gibbs law (7). *Vertical lines*: the initial distribution of money. (Reproduced from [25])

Several different rules for Δm were considered in [25]. In one model, the amount transferred was fixed to a constant $\Delta m = 1\$$. Economically, it means that all agents were selling their products for the same price $\Delta m = 1\$$. Computer animation [35] shows that the initial distribution of money first broadens to a symmetric, Gaussian curve, characteristic for a diffusion process. Then, the distribution starts to pile up around the $m = 0$ state, which acts as the impenetrable boundary, because of the imposed condition $m \geq 0$. As a result, $P(m)$ becomes skewed (asymmetric) and eventually reaches the stationary exponential shape, as shown in Fig. 1. The boundary at $m = 0$ is analogous to the ground-state energy in statistical physics. Without this boundary condition, the probability distribution of money would not reach a stationary state. Computer animation [35,36] also shows how the entropy of money distribution, defined as $S/N = -\sum_k P(m_k) \ln P(m_k)$, grows from the initial value $S = 0$, when all agents have the same money, to the maximal value at the statistical equilibrium.

While the model with $\Delta m = 1$ is very simple and instructive, it is not very realistic, because all prices are taken to be the same. In another model considered in [25], Δm in each transaction is taken to be a random fraction of the average amount of money per agent, i. e., $\Delta m = \nu(M/N)$, where ν is a uniformly distributed random number between 0 and 1. The random distribution of Δm is supposed to represent the wide variety of prices for different products in the real economy. It reflects the fact that agents buy and consume many different types of products, some of them simple and cheap, some sophisticated and expensive. Moreover, different agents like to consume these products in different quantities, so there is variation in the amounts Δm paid, even though the unit price of the same product is constant. Computer simulation of this model produces exactly the same stationary distribution (7) as in the first model. Computer animation for this model is also available on the Web page [35].

The final distribution is universal despite different rules for Δm. To amplify this point further, Drăgulescu and Yakovenko [25] also considered a toy model, where Δm was taken to be a random fraction of the average amount of money of the two agents: $\Delta m = \nu(m_i + m_j)/2$. This model produced the same stationary distribution (7) as the other two models.

The pairwise models of money transfer are attractive in their simplicity, but they represent a rather primitive market. The modern economy is dominated by big firms, which consist of many agents, so Drăgulescu and Yakovenko [25] also studied a model with firms. One agent at a time is appointed to become a "firm". The firm borrows capital K from another agent and returns it with interest hK, hires L agents and pays them wages W, manufactures Q items of a product, sells them to Q agents at price R, and receives profit $F = RQ - LW - hK$. All of these agents are randomly selected. The parameters of the model are optimized following a procedure from economics textbooks [37]. The aggregate demand–supply curve for the product is used in the form $R(Q) = v/Q^{\eta}$, where Q is the quantity consumers would buy at price R, and η and v are some parameters. The production function of the firm has the traditional Cobb–Douglas form: $Q(L, K) = L^{\chi} K^{1-\chi}$, where χ is a parameter. Then the profit of the firm F is maximized with respect to K and L. The net result of the firm activity is a many-body transfer of money, which still satisfies the conservation law. Computer simulation of this model generates the same exponential distribution (7), independently of the model parameters. The reasons for the universality of the Boltzmann–Gibbs distribution and its limitations are discussed from a different perspective in Sect. "Additive Versus Multiplicative Models".

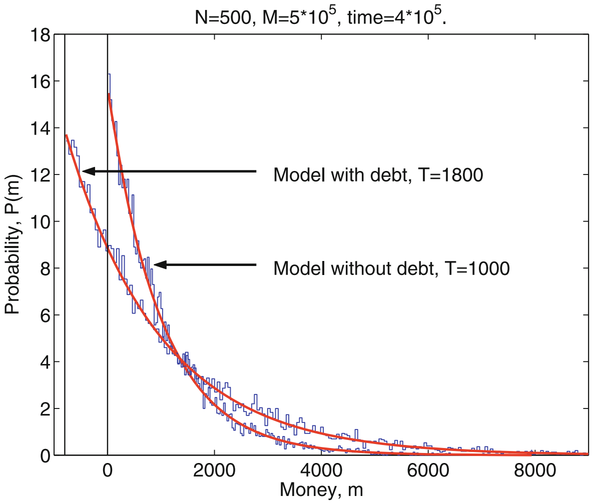
Well after paper [25] appeared, Italian econophysicists [38] found that similar ideas had been published earlier in obscure journals in Italian by Eleonora Bennati [39,40]. They proposed calling these models the Bennati–Drăgulescu–Yakovenko game [41]. The Boltzmann distribution was independently applied to social sciences by Jürgen Mimkes using the Lagrange principle of maximization with constraints [42,43]. The exponential distribution of money was also found in [44] using a Markov chain approach to strategic market games. A long time ago, Benoit Mandelbrot observed (see p. 83 in [45]):

> "There is a great temptation to consider the exchanges of money which occur in economic interaction as analogous to the exchanges of energy which occur in physical shocks between gas molecules".

He realized that this process should result in the exponential distribution, by analogy with the barometric distribution of density in the atmosphere. However, he discarded this idea, because it does not produce the Pareto power law, and proceeded to study the stable Lévy distributions. Ironically, the actual economic data, discussed in Sect. "Empirical Data on Money and Wealth Distributions" and "Empirical Data on Income Distribution", do show the exponential distribution for the majority of the population. Moreover, the data have finite variance, so the stable Lévy distributions are not applicable because of their infinite variance.

Models with Debt

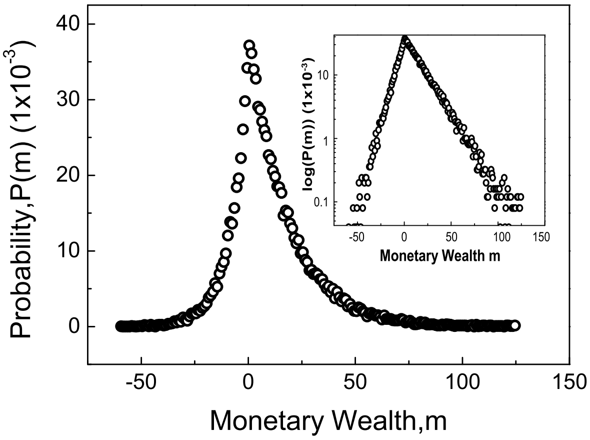
Now let us discuss how the results change when debt is permitted. Debt may be considered as negative money. When an agent borrows money from a bank (considered here as a big reservoir of money), the cash balance of the agent (positive money) increases, but the agent also acquires a debt obligation (negative money), so the total balance (net worth) of the agent remains the same, and the conservation law of total money (positive and negative) is satisfied. After spending some cash, the agent still has the debt obligation, so the money balance of the agent becomes negative. Any stable economic system must have a mechanism preventing unlimited borrowing and unlimited debt. Otherwise, agents can buy any products without producing anything in exchange by simply going into unlimited debt. The exact mechanisms of limiting debt in the real economy are complicated and obscured. Drăgulescu and Yakovenko [25] considered a simple model where the maximal debt of any agent is limited by a certain amount m_{d}. This means that the boundary condition $m_i \geq 0$ is now replaced by the condition $m_i \geq -m_{\mathrm{d}}$ for all agents i. Setting interest rates on borrowed money to be



Econophysics, Statistical Mechanics Approach to, Figure 2
Stationary distributions of money with and without debt. The debt is limited to $m_d = 800$. *Solid curves*: fits to the Boltzmann–Gibbs laws with the "temperatures" $T = 1800$ and $T = 1000$. (Reproduced from [25])

zero for simplicity, Drăgulescu and Yakovenko [25] performed computer simulations of the models described in Sect. "The Boltzmann–Gibbs Distribution of Money" with the new boundary condition. The results are shown in Fig. 2. Not surprisingly, the stationary money distribution again has an exponential shape, but now with the new boundary condition at $m = -m_d$ and the higher money temperature $T_d = m_d + M/N$. By allowing agents to go into debt up to m_d, we effectively increase the amount of money available to each agent by m_d. So, the money temperature, which is equal to the average amount of effectively available money per agent, increases. A model with nonzero interest rates was also studied in [25].

We see that debt does not violate the conservation law of money, but rather modifies boundary conditions for $P(m)$. When economics textbooks describe how "banks create money" or "debt creates money" [37], they count only positive money (cash) as money, but do not count liabilities (debt obligations) as negative money. With such a definition, money is not conserved. However, if we include debt obligations in the definition of money, then the conservation law is restored. This approach is in agreement with the principles of double-entry accounting, which records both assets and debts. Debt obligations are as real as positive cash, as many borrowers painfully realized in their experience. A more detailed study of positive and negative money and bookkeeping from the point of view of econophysics is presented in a series of papers by the physicist Dieter Braun [46,47,48].

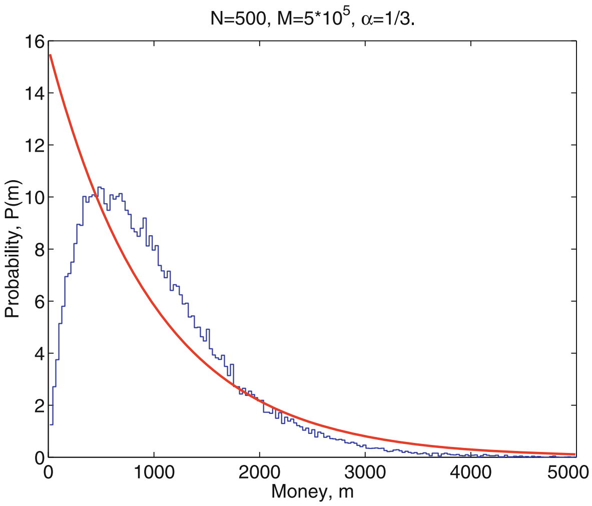


Econophysics, Statistical Mechanics Approach to, Figure 3
The stationary distribution of money for the required reserve ratio $r = 0.8$. The distribution is exponential for positive and negative money with different "temperatures" T_+ and T_-, as illustrated by the *inset* on log–linear scale. (Reproduced from [49])

One way of limiting the total debt in the economy is the so-called required reserve ratio r [37]. Every bank is required by law to set aside a fraction r of the money deposited with the bank, and this reserved money cannot be loaned further. If the initial amount of money in the system (the money base) is M_0, then with loans and borrowing the total amount of positive money available to the agents increases to $M = M_0/r$, where the factor $1/r$ is called the money multiplier [37]. This is how "banks create money". Where does this extra money come from? It comes from the increase of the total debt in the system. The maximal total debt is equal to $D = M_0/r - M_0$ and is limited by the factor r. When the debt is maximal, the total amounts of positive, M_0/r, and negative, $M_0(1-r)/r$, money circulate between the agents in the system, so there are effectively two conservation laws for each of them [49]. Thus, we expect to see the exponential distributions of positive and negative money characterized by two different temperatures: $T_+ = M_0/rN$ and $T_- = M_0(1-r)/rN$. This is exactly what was found in computer simulations in [49], shown in Fig. 3. Similar two-sided distributions were also found in [47].

Proportional Money Transfers and Saving Propensity

In the models of money transfer considered thus far, the transferred amount Δm is typically independent of the money balances of agents. A different model was introduced in the physics literature earlier [50] under the name multiplicative asset exchange model. This model also satisfies the conservation law, but the amount of money trans-



Econophysics, Statistical Mechanics Approach to, Figure 4
Stationary probability distribution of money in the multiplicative random exchange model (8) for $\gamma = 1/3$. *Solid curve*: the exponential Boltzmann–Gibbs law. (Reproduced from [25])

ferred is a fixed fraction γ of the payer's money in (5):

$$\Delta m = \gamma m_i \, . \tag{8}$$

The stationary distribution of money in this model, shown in Fig. 4 with an exponential function, is close, but not exactly equal, to the Gamma distribution:

$$P(m) = c m^{\beta} \mathrm{e}^{\frac{-m}{T}} \, . \tag{9}$$

Equation (9) differs from (7) by the power-law prefactor m^{β}. From the Boltzmann kinetic equation (discussed in more detail in Sect. "Additive Versus Multiplicative Models"), Ispolatov et al. [50] derived a formula relating the parameters γ and β in (8) and (9):

$$\beta = \frac{-1 - \ln 2}{\ln(1 - \gamma)} \, . \tag{10}$$

When payers spend a relatively small fraction of their money $\gamma < 1/2$, (10) gives $\beta > 0$, so the low-money population is reduced and $P(m \to 0) = 0$, as shown in Fig. 4.

Later, Thomas Lux brought to the attention of physicists [32] that essentially the same model, called the inequality process, had been introduced and studied much earlier by the sociologist John Angle [51,52,53,54,55], see also the review [56] for additional references. While Ispolatov et al. [50] did not give much justification for the proportionality law (8), Angle [51] connected this rule with the surplus theory of social stratification [57], which argues that inequality in human society develops when people can produce more than necessary for minimal subsistence. This additional wealth (surplus) can be transferred from original producers to other people, thus generating inequality. In the first paper by Angle [51], the parameter γ was randomly distributed, and another parameter, δ, gave a higher probability of winning to the agent with a higher money balance in (5). However, in the following papers, he simplified the model to a fixed γ (denoted as ω by Angle) and equal probabilities of winning for higher- and lower-balance agents, which makes it completely equivalent to the model of [50]. Angle also considered a model [55,56] where groups of agents have different values of γ, simulating the effect of education and other "human capital". All of these models generate a Gamma-like distribution, well approximated by (9).

Another model with an element of proportionality was proposed in [26]. (This paper originally appeared as a follow-up preprint cond-mat/0004256 to the preprint cond-mat/0001432 of [25].) In this model, the agents set aside (save) some fraction of their money λm_i, whereas the rest of their money balance $(1 - \lambda) m_i$ becomes available for random exchanges. Thus, the rule of exchange (5) becomes

$$\begin{aligned} m_i' &= \lambda m_i + \xi (1 - \lambda)(m_i + m_j) \, , \\ m_j' &= \lambda m_j + (1 - \xi)(1 - \lambda)(m_i + m_j) \, . \end{aligned} \tag{11}$$

Here the coefficient λ is called the saving propensity, and the random variable ξ is uniformly distributed between 0 and 1. It was pointed out in [56] that, by the change of notation $\lambda \to (1 - \gamma)$, (11) can be transformed to the same form as (8), if the random variable ξ takes only discrete values 0 and 1. Computer simulations [26] of the model (11) found a stationary distribution close to the Gamma distribution (9). It was shown that the parameter β is related to the saving propensity λ by the formula $\beta = 3\lambda/(1 - \lambda)$ [38,58,59,60]. For $\lambda \neq 0$, agents always keep some money, so their balances never go to zero and $P(m \to 0) = 0$, whereas for $\lambda = 0$ the distribution becomes exponential.

In the subsequent papers by the Kolkata school [1] and related papers, the case of random saving propensity was studied. In these models, the agents are assigned random parameters λ drawn from a uniform distribution between 0 and 1 [61]. It was found that this model produces a power-law tail $P(m) \propto 1/m^2$ at high m. The reasons for stability of this law were understood using the Boltzmann kinetic equation [60,62,63], but most elegantly in the mean-field theory [64]. The fat tail originates from the agents whose saving propensity is close to 1, who hoard money and do not give it back [38,65]. An interesting matrix formulation of the problem was presented in [66].

Patriarca et al. [67] studied the relaxation rate in the money transfer models. Drăgulescu and Yakovenko [25] studied a model with taxation, which also has an element of proportionality. The Gamma distribution was also studied for conservative models within a simple Boltzmann approach in [68] and using much more complicated rules of exchange in [69,70].

Additive Versus Multiplicative Models

The stationary distribution of money (9) for the models of Sect. "Proportional Money Transfers and Saving Propensity" is different from the simple exponential formula (7) found for the models of Sect. "The Boltzmann–Gibbs Distribution of Money". The origin of this difference can be understood from the Boltzmann kinetic equation [28,71]. This equation describes time evolution of the distribution function $P(m)$ due to pairwise interactions:

$$\frac{\mathrm{d}P(m)}{\mathrm{d}t} = \iint \{-f_{[m,m']\to[m-\Delta,m'+\Delta]}P(m)P(m') + f_{[m-\Delta,m'+\Delta]\to[m,m']}P(m-\Delta)\cdot P(m'+\Delta)\}\,\mathrm{d}m'\mathrm{d}\Delta\ . \tag{12}$$

Here $f_{[m,m']\to[m-\Delta,m'+\Delta]}$ is the probability of transferring money Δ from an agent with money m to an agent with money m' per unit time. This probability, multiplied by the occupation numbers $P(m)$ and $P(m')$, gives the rate of transitions from the state $[m, m']$ to the state $[m-\Delta, m'+\Delta]$. The first term in (12) gives the depopulation rate of the state m. The second term in (12) describes the reverse process, where the occupation number $P(m)$ increases. When the two terms are equal, the direct and reverse transitions cancel each other statistically, and the probability distribution is stationary: $\mathrm{d}P(m)/\mathrm{d}t = 0$. This is the principle of detailed balance.

In physics, the fundamental microscopic equations of motion of particles obey time-reversal symmetry. This means that the probabilities of the direct and reverse processes are exactly equal:

$$f_{[m,m']\to[m-\Delta,m'+\Delta]} = f_{[m-\Delta,m'+\Delta]\to[m,m']}\ . \tag{13}$$

When (13) is satisfied, the detailed balance condition for (12) reduces to the equation $P(m)P(m') = P(m-\Delta)P(m'+\Delta)$, because the factors f cancel out. The only solution of this equation is the exponential function $P(m) = c\exp(-m/T_m)$, so the Boltzmann–Gibbs distribution is the stationary solution of the Boltzmann kinetic equation (12). Notice that the transition probabilities (13) are determined by the dynamical rules of the model, but the equilibrium Boltzmann–Gibbs distribution does not depend on the dynamical rules at all. This is the origin of the universality of the Boltzmann–Gibbs distribution. It shows that it may be possible to find out the stationary distribution without knowing details of the dynamical rules (which are rarely known very well), as long as the symmetry condition (13) is satisfied.

The models considered in Sect. "The Boltzmann–Gibbs Distribution of Money" have the time-reversal symmetry. The model with the fixed money transfer Δ has equal probabilities (13) of transferring money from an agent with balance m to an agent with balance m' and vice versa. This is also true when Δ is random, as long as the probability distribution of Δ is independent of m and m'. Thus, the stationary distribution $P(m)$ is always exponential in these models.

However, there is no fundamental reason to expect time-reversal symmetry in economics, so (13) may be not valid. In this case, the system may have a nonexponential stationary distribution or no stationary distribution at all. In model (8), the time-reversal symmetry is broken. Indeed, when an agent i gives a fixed fraction γ of his money m_i to an agent with balance m_j, their balances become $(1-\gamma)m_i$ and $m_j + \gamma m_i$. If we try to reverse this process and appoint an agent j to be the payer and to give the fraction γ of her money, $\gamma(m_j + \gamma m_i)$, to agent i, the system does not return to the original configuration $[m_i, m_j]$. As emphasized by Angle [56], the payer pays a deterministic fraction of his money, but the receiver receives a random amount from a random agent, so their roles are not interchangeable. Because the proportional rule typically violates the time-reversal symmetry, the stationary distribution $P(m)$ in multiplicative models is typically not exactly exponential.[1]Making the transfer dependent on the money balance of the payer effectively introduces a Maxwell's demon into the model. That is why the stationary distribution is not exponential, and, thus, does not maximize entropy (4). Another view on the time-reversal symmetry in economic dynamics is presented in [72].

These examples show that the Boltzmann–Gibbs distribution does not hold for any conservative model. However, it is universal in a limited sense. For a broad class of models that have time-reversal symmetry, the stationary distribution is exponential and does not depend on the details of the model. Conversely, when the time-reversal symmetry is broken, the distribution may depend on the details of the model. The difference between these two

[1] However, when Δm is a fraction of the total money $m_i + m_j$ of the two agents, the model is time-reversible and has an exponential distribution, as discussed in Sect. "The Boltzmann–Gibbs Distribution of Money".

classes of models may be rather subtle. Deviations from the Boltzmann–Gibbs law may occur only if the transition rates f in (13) explicitly depend on the agent's money m or m' in an asymmetric manner. Drăgulescu and Yakovenko [25] performed a computer simulation where the direction of payment was randomly selected in advance for every pair of agents (i, j). In this case, money flows along directed links between the agents: $i \to j \to k$, and the time-reversal symmetry is strongly violated. This model is closer to the real economy, where one typically receives money from an employer and pays it to a grocery store. Nevertheless, the Boltzmann–Gibbs distribution was found in this model, because the transition rates f do not explicitly depend on m and m' and do not violate (13).

In the absence of detailed knowledge of real microscopic dynamics of economic exchanges, the semiuniversal Boltzmann–Gibbs distribution (7) is a natural starting point. Moreover, the assumption of [25] that agents pay the same prices Δm for the same products, independent of their money balances m, seems very appropriate for the modern anonymous economy, especially for purchases over the Internet. There is no particular empirical evidence for the proportional rules (8) or (11). However, the difference between the additive (7) and multiplicative (9) distributions may be not so crucial after all. From the mathematical point of view, the difference is in the implementation of the boundary condition at $m = 0$. In the additive models of Sect. "The Boltzmann–Gibbs Distribution of Money", there is a sharp cutoff of $P(m)$ at $m = 0$. In the multiplicative models of Sect. "Proportional Money Transfers and Saving Propensity", the balance of an agent never reaches $m = 0$, so $P(m)$ vanishes at $m \to 0$ in a power-law manner. At the same time, $P(m)$ decreases exponentially for large m for both models.

By further modifying the rules of money transfer and introducing more parameters in the models, one can obtain even more complicated distributions [73]. However, one can argue that parsimony is the virtue of a good mathematical model, not the abundance of additional assumptions and parameters, whose correspondence to reality is hard to verify.

Statistical Mechanics of Wealth Distribution

In the econophysics literature on exchange models, the terms "money" and "wealth" are often used interchangeably; however, economists emphasize the difference between these two concepts. In this section, we review the models of wealth distribution, as opposed to money distribution.

Models with a Conserved Commodity

What is the difference between money and wealth? One can argue [25] that wealth w_i is equal to money m_i plus the other property that an agent i has. The latter may include durable material property, such as houses and cars, and financial instruments, such as stocks, bonds, and options. Money (paper cash, bank accounts) is generally liquid and countable. However, the other property is not immediately liquid and has to be sold first (converted into money) to be used for other purchases. In order to estimate the monetary value of property, one needs to know the price p. In the simplest model, let us consider just one type of property, say, stocks s. Then the wealth of an agent i is given by the formula

$$w_i = m_i + p s_i \ . \tag{14}$$

It is assumed that the price p is common for all agents and is established by some kind of market process, such as an auction, and may change in time.

It is reasonable to start with a model where both the total money $M = \sum_i m_i$ and the total stock $S = \sum_i s_i$ are conserved [74,75,76]. The agents pay money to buy stock and sell stock to get money, and so on. Although M and S are conserved, the total wealth $W = \sum_i w_i$ is generally not conserved, because of the price fluctuation [75] in (14). This is an important difference from the money transfer models of Sect. "Statistical Mechanics of Money Distribution". Here the wealth w_i of an agent i, not participating in any transactions, may change when transactions between other agents establish a new price p. Moreover, the wealth w_i of an agent i does not change after a transaction with an agent j. Indeed, in exchange for paying money Δm, agent i receives the stock $\Delta s = \Delta m/p$, so her total wealth (14) remains the same. In principle, the agent can instantaneously sell the stock back at the same price and recover the money paid. If the price p never changes, then the wealth w_i of each agent remains constant, despite transfers of money and stock betweenagents.

We see that redistribution of wealth in this model is directly related to price fluctuations. One mathematical model of this process was studied in [77]. In this model, the agents randomly change preferences for the fraction of their wealth invested in stocks. As a result, some agents offer stock for sale and some want to buy it. The price p is determined from the market-clearing auction matching supply and demand. Silver et al. [77] demonstrated in computer simulations and proved analytically using the theory of Markov processes that the stationary distribution $P(w)$ of wealth w in this model is given by the Gamma distribution, as in (9). Various modifications of this model [32], such as introducing monopolistic coalitions, do not change this result significantly, which shows the robustness of the Gamma distribution. For models with a conserved commodity, Chatterjee and Chakrabarti [75] found the Gamma distribution for a fixed saving propensity and a power law tail for a distributed saving propensity.

Another model with conserved money and stock was studied in [78] for an artificial stock market where traders follow different investment strategies: random, momentum, contrarian, and fundamentalist. Wealth distribution in the model with random traders was found have a power-law tail $P(w) \sim 1/w^2$ for large w. However, unlike in most other simulation, where all agents initially have equal balances, here the initial money and stock balances of the agents were randomly populated according to a power law with the same exponent. This raises the question whether the observed power-law distribution of wealth is an artifact of the initial conditions, because equilibrization of the upper tail may take a very long simulation time.

Models with Stochastic Growth of Wealth

Although the total wealth W is not exactly conserved in the models considered in Sect. "Models with a Conserved Commodity", W nevertheless remains constant on average, because the total money M and stock S are conserved. A different model for wealth distribution was proposed in [27]. In this model, time evolution of the wealth w_i of an agent i is given by the stochastic differential equation

$$\frac{\mathrm{d}w_i}{\mathrm{d}t} = \eta_i(t)w_i + \sum_{j(\neq i)} J_{ij}w_j - \sum_{j(\neq i)} J_{ji}w_i \,, \tag{15}$$

where $\eta_i(t)$ is a Gaussian random variable with mean $\langle\eta\rangle$ and variance $2\sigma^2$. This variable represents growth or loss of wealth of an agent due to investment in stock market. The last two terms describe transfer of wealth between different agents, which is taken to be proportional to the wealth of the payers with the coefficients J_{ij}. So, the model (15) is multiplicative and invariant under the scale transformation $w_i \to Zw_i$. For simplicity, the exchange fractions are taken to be the same for all agents: $J_{ij} = J/N$ for all $i \neq j$, where N is the total number of agents. In this case, the last two terms in (15) can be written as $J(\langle w\rangle - w_i)$, where $\langle w\rangle = \sum_i w_i/N$ is the average wealth per agent. This case represents a "mean-field" model, where all agents feel the same environment. It can be easily shown that the average wealth increases in time as $\langle w\rangle_t = \langle w\rangle_0 \mathrm{e}^{(\langle\eta\rangle+\sigma^2)t}$. Then, it makes more sense to consider the relative wealth $\tilde{w}_i = w_i/\langle w\rangle_t$. Equation (15) for this variable becomes

$$\frac{\mathrm{d}\tilde{w}_i}{\mathrm{d}t} = (\eta_i(t) - \langle\eta\rangle - \sigma^2)\tilde{w}_i + J(1-\tilde{w}_i)\,. \tag{16}$$

The probability distribution $P(\tilde{w},t)$ for the stochastic differential equation (16) is governed by the Fokker–Planck equation:

$$\frac{\partial P}{\partial t} = \frac{\partial[J(\tilde{w}-1)+\sigma^2\tilde{w}]P}{\partial\tilde{w}} + \sigma^2\frac{\partial}{\partial\tilde{w}}\left(\tilde{w}\frac{\partial(\tilde{w}P)}{\partial\tilde{w}}\right). \tag{17}$$

The stationary solution ($\partial P/\partial t = 0$) of this equation is given by the following formula:

$$P(\tilde{w}) = c\frac{\mathrm{e}^{\frac{-J}{\sigma^2\tilde{w}}}}{\tilde{w}^{\frac{2+J}{\sigma^2}}}\,. \tag{18}$$

The distribution (18) is quite different from the Boltzmann–Gibbs (7) and Gamma (9) distributions. Equation (18) has a power-law tail at large $\tilde{w}$ and a sharp cutoff at small $\tilde{w}$. Equation (15) is a version of the generalized Lotka–Volterra model, and the stationary distribution (18) was also obtained in [79,80]. The model was generalized to include negative wealth in [81].

Bouchaud and Mézard [27] used the mean-field approach. A similar result was found for a model with pairwise interaction between agents in [82]. In this model, wealth is transferred between the agents using the proportional rule (8). In addition, the wealth of the agents increases by the factor $1+\zeta$ in each transaction. This factor is supposed to reflect creation of wealth in economic interactions. Because the total wealth in the system increases, it makes sense to consider the distribution of relative wealth $P(\tilde{w})$. In the limit of continuous trading, Slanina [82] found the same stationary distribution (18). This result was reproduced using a mathematically more involved treatment of this model in [83]. Numerical simulations of the models with stochastic noise η in [69,70] also found a power-law tail for large w.

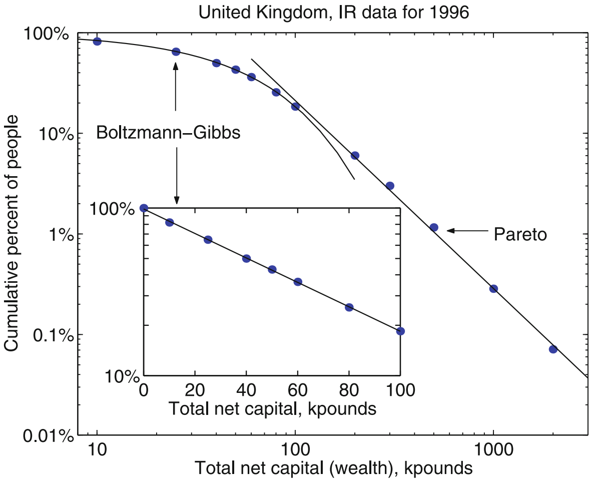
Let us contrast the models discussed in Sect. "Models with a Conserved Commodity" and "Models with Stochastic Growth of Wealth". In the former case, where money and commodity are conserved and wealth does not grow, the distribution of wealth is given by the Gamma distribution with an exponential tail for large w. In the latter models, wealth grows in time exponentially, and the distribution of relative wealth has a power-law tail for large $\tilde{w}$. These results suggest that the presence of a power-law tail is a nonequilibrium effect that requires constant growth or inflation of the economy, but disappears for a closed system with conservation laws.

Reviews of the models discussed were also given in [84,85]. Because of lack of space, we omit discussion of models with wealth condensation [27,50,86,87,88], where a few agents accumulate a finite fraction of the total wealth, and studies of wealth distribution on networks [89,90,91,92]. Here we discuss the models with long-range interaction, where any agent can exchange money and wealth with any other agent. A local model, where agents trade only with the nearest neighbors, was studied in [93].

Empirical Data on Money and Wealth Distributions

It would be very interesting to compare theoretical results for money and wealth distributions in various models with empirical data. Unfortunately, such empirical data are difficult to find. Unlike income, which is discussed in Sect. "Data and Models for Income Distribution", wealth is not routinely reported by the majority of individuals to the government. However, in many countries, when a person dies, all assets must be reported for the purpose of inheritance tax. So, in principle, there exist good statistics of wealth distribution among dead people, which, of course, is different from the wealth distribution among living people. Using an adjustment procedure based on the age, gender, and other characteristics of the deceased, the UK tax agency, the Inland Revenue, reconstructed the wealth distribution of the whole population of the UK [94]. Figure 5 shows the UK data for 1996 reproduced from [95]. The figure shows the cumulative probability $C(w) = \int_w^\infty P(w')\mathrm{d}w'$ as a function of the personal net wealth w, which is composed of assets (cash, stocks, property, household goods, etc.) and liabilities (mortgages and other debts). Because statistical data are usually reported at nonuniform intervals of w, it is more practical to plot the cumulative probability distribution $C(w)$ rather than its derivative, the probability density $P(w)$. Fortunately, when $P(w)$ is an exponential or a power-law function, then $C(w)$ is also an exponential or a power-law function.

The cumulative probability distribution in Fig. 5 is plotted on a log–log scale, where a straight line represents a power-law dependence. The figure shows that the distribution follows a power law $C(w) \propto 1/w^\alpha$ with exponent $\alpha = 1.9$ for wealth greater than about £100,000. The inset in Fig. 5 shows the data on log–linear scale, where the straight line represents an exponential dependence. We observe that below £100,000 the data are well fitted by the exponential distribution $C(w) \propto \exp(-w/T_w)$ with the effective "wealth temperature" $T_w = £60{,}000$, (which corresponds to the median wealth of £41,000). So, the distribution of wealth is characterized by the Pareto power law in the upper tail of the distribution and the exponential Boltzmann–Gibbs law in the lower part of the distribution for the great majority (about 90%) of the population. Similar results are found for the distribution of income, as discussed in Sect. "Data and Models for Income Distribution". One may speculate that the wealth distribution in the lower part is dominated by distribution of money, because the corresponding people do not have other significant assets, so the results of Sect. "Statistical Mechanics of Money Distribution" give the Boltzmann–Gibbs law. On the other hand, the upper tail of the wealth distribution is dominated by investment assess, where the results of Sect. "Models with Stochastic Growth of Wealth" give the Pareto law. The power law was studied by many researchers for the upper-tail data, such as the Forbes list of the 400 richest people [96,97], but much less attention was paid to the lower part of the wealth distribution. Curiously, Abdul-Magd [98] found that the wealth distribution in ancient Egyptian society was consistent with (18).



Econophysics, Statistical Mechanics Approach to, Figure 5
Cumulative probability distribution of net wealth in the UK shown on log–log and log–linear (*inset*) scales. *Points* represent the data from the Inland Revenue, and *solid lines* are fits to the exponential (Boltzmann–Gibbs) and power (Pareto) laws. (Reproduced from [95])

For direct comparison with the results of Sect. "Statistical Mechanics of Money Distribution", it would be very interesting to find data on the distribution of money, as opposed to the distribution of wealth. Making a reasonable assumption that most people keep most of their money in banks, one can approximate the distribution of money by the distribution of balances on bank accounts. (Balances on all types of bank accounts, such as checking, saving, and money manager, associated with the same person should be added up.) Despite imperfections (people may have accounts with different banks or not keep all their money in banks), the distribution of balances on bank accounts would give valuable information about the distribution of money. The data for a big enough bank would be representative of the distribution in the whole economy. Unfortunately, it has not been possible to obtain such data thus far, even though it would be completely anonymous and not compromise the privacy of bank clients.

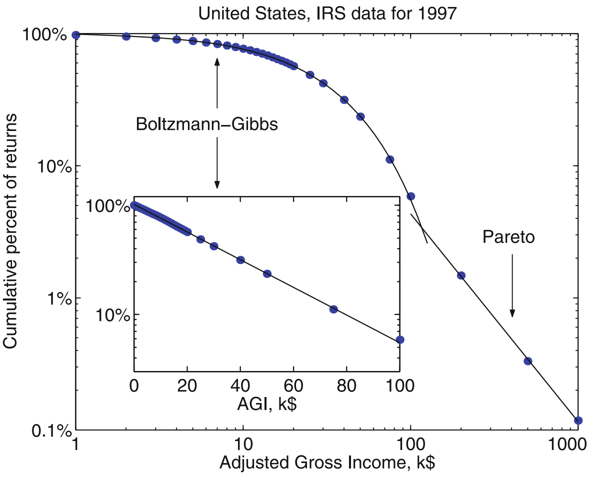
Measuring the probability distribution of money would be very useful, because it determines how much people can, in principle, spend on purchases without going into debt. This is different from the distribution of wealth, where the property component, such as house, car, or retirement investment, is effectively locked up and, in most cases, is not easily available for consumer spending. So, although wealth distribution may reflect the distribution of economic power, the distribution of money is more relevant for consumption. Money distribution can be useful for determining prices that maximize revenue of a manufacturer [25]. If a price p is set too high, few people can afford it, and, if a price is too low, the sales revenue is small, so the optimal price must be in-between. The fraction of the population who can afford to pay the price p is given by the cumulative probability $C(p)$, so the total sales revenue is proportional to $pC(p)$. For the exponential distribution $C(p) = \exp(-p/T_m)$, the maximal revenue is achieved at $p = T_m$, i. e., at the optimal price equal to the average amount of money per person [25]. Indeed, the prices of mass-market consumer products, such as computers, electronics goods, and appliances, remain stable for many years at a level determined by their affordability to the population, whereas the technical parameters of these products at the same price level improve dramatically owing to technological progress.

Data and Models for Income Distribution

In contrast to money and wealth distributions, a lot more empirical data are available for the distribution of income r from tax agencies and population surveys. In this section, we first present empirical data on income distribution and then discuss theoretical models.

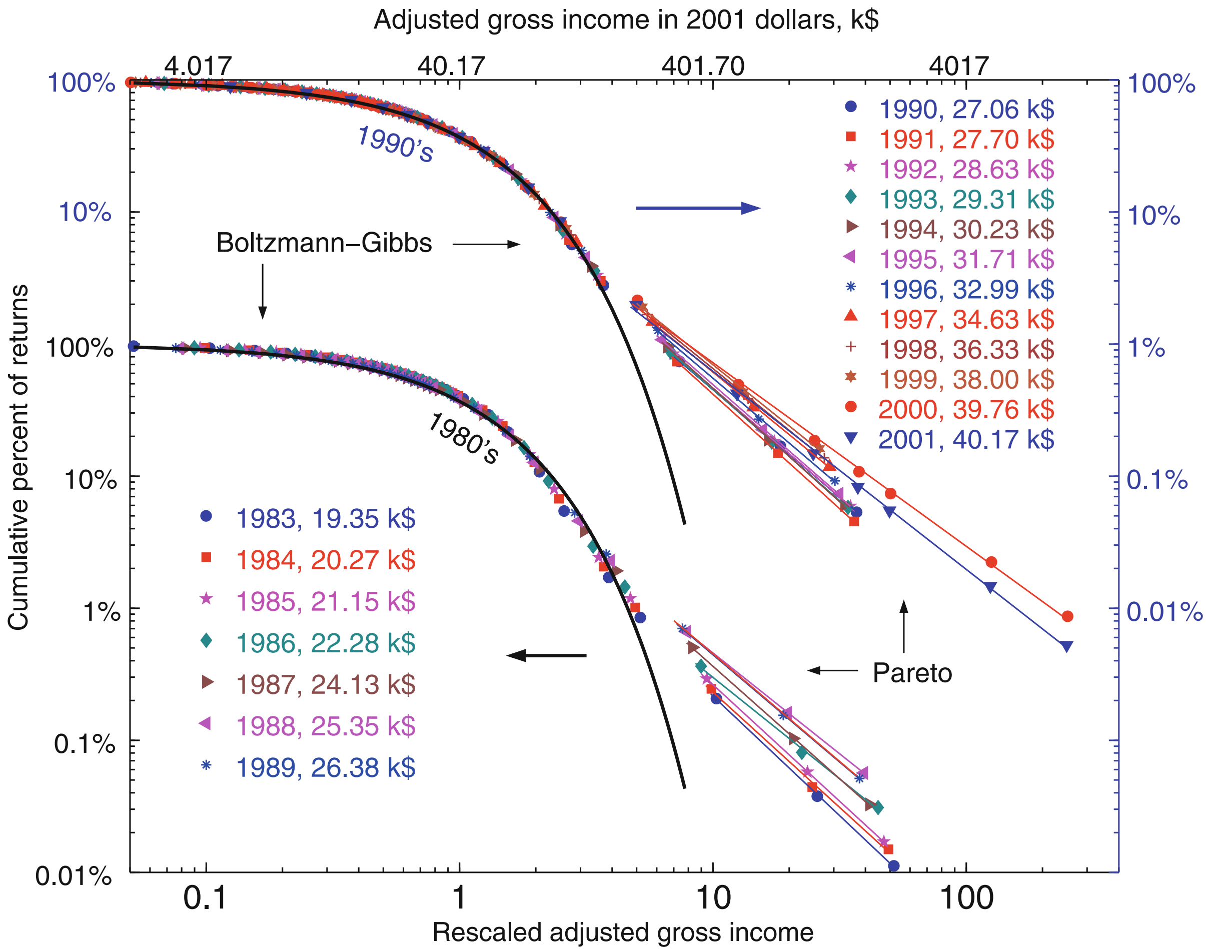
Empirical Data on Income Distribution

Empirical studies of income distribution have a long history in the economics literature [99,100,101]. Following the work by Pareto [15], much attention was focused on the power-law upper tail of the income distribution and less on the lower part. In contrast to more complicated functions discussed in the literature, Drăgulescu and Yakovenko [102] introduced a new idea by demonstrating that the lower part of income distribution can be well fitted with a simple exponential function $P(r) = c\exp(-r/T_r)$ characterized by just one parameter, the "income temperature" T_r. Then it was recognized that the whole income distribution can be fitted by an exponential function in the lower part and a power-law function in the upper part [95,103], as shown in Fig. 6. The straight line on the log–linear scale in the inset of Fig. 6 demonstrates the exponential Boltzmann–Gibbs law, and the straight line on the log–log scale in the main panel illustrates the Pareto power law. The fact that income distribution consists of two distinct parts reveals the two-class structure of American society [104,105]. Coexistence of the exponential and power-law distributions is known in plasma physics and astrophysics, where they are called the "thermal" and "superthermal"parts [106,107,108]. The boundary between the lower and upper classes can be defined as the intersection point of the exponential and power-law fits in Fig. 6. For 1997, the annual income separating the two classes was about $120,000. About 3% of the population belonged to the upper class, and 97% belonged to the lower class.



Econophysics, Statistical Mechanics Approach to, Figure 6
Cumulative probability distribution of tax returns for USA in 1997 shown on log–log and log–linear (*inset*) scales. *Points* represent the Internal Revenue Service (IRS) data, and *solid lines* are fits to the exponential and power-law functions. (Reproduced from [103])

Silva and Yakovenko [105] studied time evolution of income distribution in the USA during 1983–2001 on the basis of data from the Internal Revenue Service (IRS), the government tax agency. The structure of the income distribution was found to be qualitatively the same for all years, as shown in Fig. 7. The average income in nominal dollars approximately doubled during this time interval. So,



Econophysics, Statistical Mechanics Approach to, Figure 7
Cumulative probability distribution of tax returns plotted on log–log scale versus r/T_r (the annual income r normalized by the average income T_r in the exponential part of the distribution). The IRS data points are for 1983–2001, and the *columns of numbers* give the values of T_r for the corresponding years. (Reproduced from [105])

the horizonal axis in Fig. 7 shows the normalized income r/T_r, where the"income temperature" T_r was obtained by fitting of the exponential part of the distribution for each year. The values of T_r are shown in Fig. 7. The plots for the 1980s and 1990s are shifted vertically for clarity. We observe that the data points in the lower-income part of the distribution collapse on the same exponential curve for all years. This demonstrates that the shape of the income distribution for the lower class is extremely stable and does not change in time, despite the gradual increase of the average income in nominal dollars. This observation suggests that the lower-class distribution is in statistical, "thermal" equilibrium.

On the other hand, Fig. 7 shows that the income distribution in the upper class does not rescale and significantly changes in time. Silva and Yakovenko [105] found that the exponent α of the power law $C(r) \propto 1/r^{\alpha}$ decreased from 1.8 in 1983 to 1.4 in 2000. This means that the upper tail became "fatter". Another useful parameter is the total income of the upper class as the fraction f of the total income in the system. The fraction f increased from 4% in 1983 to 20% in 2000 [105]. However, in 2001, α increased and f decreased, indicating that the upper tail was reduced after the stock market crash at that time. These results indicate that the upper tail is highly dynamical and not stationary. It tends to swell during the stock market boom and shrink during the bust. Similar results were found for Japan [109,110,111,112].

Although relative income inequality within the lower class remains stable, the overall income inequality in the

USA has increased significantly as a result of the tremendous growth of the income of the upper class. This is illustrated by the Lorenz curve and the Gini coefficient shown in Fig. 8. The Lorenz curve [99] is a standard way of representing income distribution in the economics literature. It is defined in terms of two coordinates $x(r)$ and $y(r)$ depending on a parameter r:

$$\begin{aligned} x(r) &= \int_0^r P(r')\mathrm{d}r' \,, \\ y(r) &= \frac{\int_0^r r'P(r')\mathrm{d}r'}{\int_0^\infty r'P(r')\mathrm{d}r'} \,. \end{aligned} \tag{19}$$

The horizontal coordinate $x(r)$ is the fraction of the population with income below r, and the vertical coordinate $y(r)$ is the fraction of the income this population accounts for. As r changes from 0 to ∞, x and y change from 0 to 1 and parametrically define a curve in the (x, y)-plane.

Figure 8 shows the data points for the Lorenz curves in 1983 and 2000, as computed by the IRS [113]. Drăgulescu and Yakovenko [102] analytically derived the Lorenz curve formula $y = x + (1-x)\ln(1-x)$ for a purely exponential distribution $P(r) = c\exp(-r/T_r)$. This formula is shown by the red line in Fig. 8 and describes the 1983 data reasonably well. However, for 2000, it is essential to take into account the fraction f of income in the upper tail, which modifies the Lorenz formula as follows [103,104,105]:

$$y = (1-f)[x + (1-x)\ln(1-x)] + f\Theta(x-1)\,. \tag{20}$$

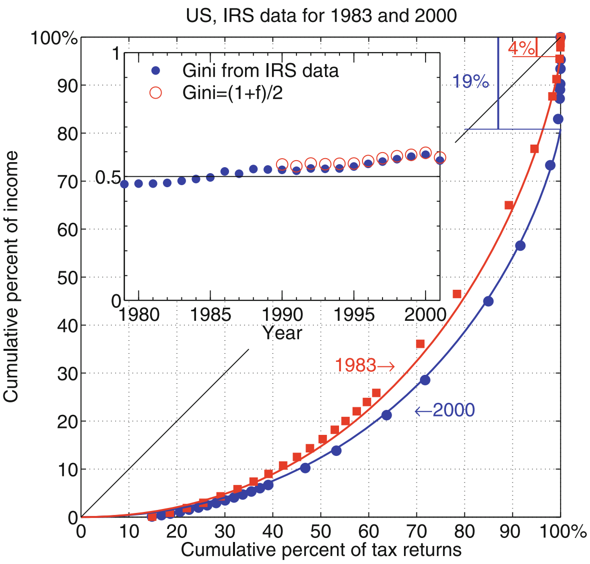
The last term in (20) represent the vertical jump of the Lorenz curve at $x = 1$, where a very small percentage of the population in the upper class accounts for a substantial fraction f of the total income. The blue curve representing (20) fits the 2000 data in Fig. 8 very well.

The deviation of the Lorenz curve from the straight diagonal line in Fig. 8 is a certain measure of income inequality. Indeed, if everybody had the same income, the Lorenz curve would be a diagonal line, because the fraction of income would be proportional to the fraction of the population. The standard measure of income inequality is the so-called Gini coefficient $0 \le G \le 1$, which is defined as the area between the Lorenz curve and the diagonal line, divided by the area of the triangle beneath the diagonal line [99]. Time evolution of the Gini coefficient, as computed by the IRS [113], is shown in the inset of Fig. 8. Drăgulescu and Yakovenko [102] derived analytically the result that $G = 1/2$ for a purely exponential distribution. In the first approximation, the values of G shown in the inset of Fig. 8 are indeed close to the theoretical value 1/2. If we take into account the upper tail using (20), the formula for the Gini coefficient becomes $G = (1+f)/2$ [105]. The inset in Fig. 8 shows that this formula is a very good fit to the IRS data for the 1990s using the values of f deduced from Fig. 7. The values $G < 1/2$ for the 1980s cannot be captured by this formula, because the Lorenz data points are slightly above the theoretical curve for 1983 in Fig. 8. Overall, we observe that income inequality has been increasing for the last 20 years, because of swelling of the Pareto tail, but decreased in 2001 after the stock market crash.

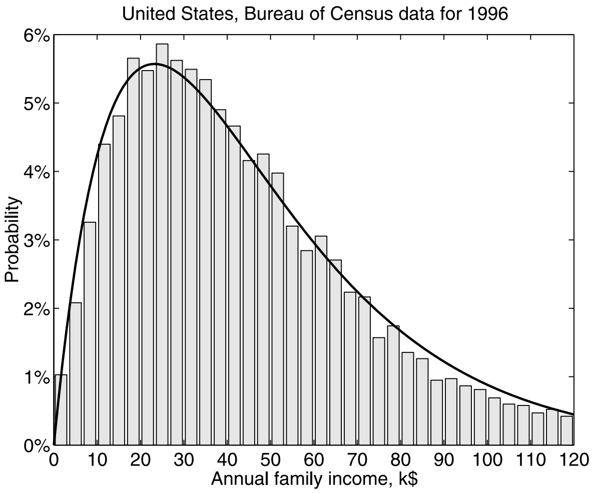
Thus far we have discussed the distribution of individual income. An interesting related question is the distribution $P_2(r)$ of family income $r = r_1 + r_2$, where r_1 and r_2 are the incomes of spouses. If individual incomes are distributed exponentially $P(r) \propto \exp(-r/T_r)$, then

$$P_2(r) = \int_0^r \mathrm{d}r' P(r')P(r-r') = cr\exp(-r/T_r)\,, \tag{21}$$

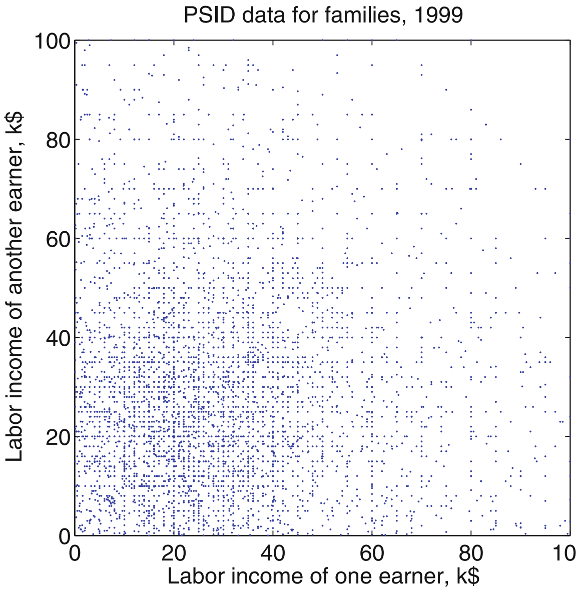
where c is a normalization constant. Figure 9 shows that (21) is in good agreement with the family income distribution data from the US Census Bureau [102]. In (21), we assumed that incomes of spouses are uncorrelated. This simple approximation is indeed supported by the scatter plot of incomes of spouses shown in Fig. 10. Each family is represented in this plot by two points (r_1, r_2) and (r_2, r_1) for symmetry. We observe that the density of points is approximately constant along the lines of constant family in-



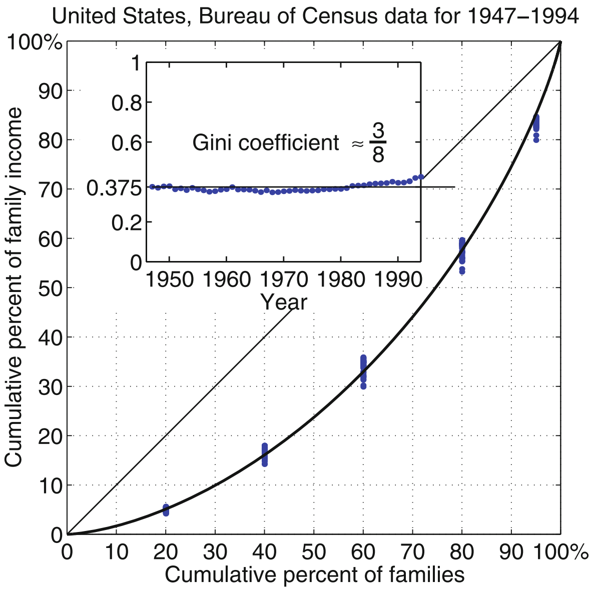
Econophysics, Statistical Mechanics Approach to, Figure 8
Lorenz plots for income distribution in 1983 and 2000. The data points are from the IRS [113], and the theoretical curves represent (20) with *f* from Fig. 7. *Inset*: The *closed circles* are the IRS data 113 for the Gini coefficient *G*, and the *open circles* show the theoretical formula *G* = (1 + *f*)/2. (Reproduced from [105])



Econophysics, Statistical Mechanics Approach to, Figure 9
Probability distribution of family income for families with two adults (US Census Bureau data). *Solid line*: fit to (21). (Reproduced from [102])



Econophysics, Statistical Mechanics Approach to, Figure 10
Scatter plot of the spouses' incomes (r_1, r_2) and (r_2, r_1) based on the data from the Panel Study of Income Dynamics (PSID). (Reproduced from [103])



Econophysics, Statistical Mechanics Approach to, Figure 11
Lorenz plot for family income calculated from (21), compared with the US Census data points. *Inset*: The US Census data points for the Gini coefficient for families, compared with the theoretically calculated value 3/8=37.5%. (Reproduced from [102])

come $r_1 + r_2 = \text{const}$, which indicates that incomes of spouses are approximately uncorrelated. There is no significant clustering of points along the diagonal $r_1 = r_2$, i. e., no strong positive correlation of spouses' incomes.

The Gini coefficient for the family income distribution (21) was calculated in [102] as $G = 3/8 = 37.5\%$. Figure 11 shows the Lorenz quintiles and the Gini coefficient for 1947–1994 plotted from the US Census Bureau data. The solid line, representing the Lorenz curve calculated from (21), is in good agreement with the data. The systematic deviation for the top 5% of earners results from the upper tail, which has a less pronounced effect on family income than on individual income, because of income averaging in the family. The Gini coefficient, shown in the inset of Fig. 11, is close to the calculated value of 37.5%. Moreover, the average G for the developed capitalist countries of North America and western Europe, as determined by the World Bank [103], is also close to the calculated value of 37.5%.

Income distribution has been examined in econophysics papers for different countries: Japan [68,109,110,111,112,114,115,116], Germany [117,118], the UK [68,85,116,117,118], Italy [118,119,120], the USA [117,121], India [97], Australia [91,120,122], and New Zealand [68,116]. The distributions are qualitatively similar to the results presented in this section. The upper tail follows a power law and comprises a small fraction of the population. To fit the lower part of the distribution, the use of exponential, Gamma, and log-normal distributions was reported in different papers. Unfortunately, income distribution is often reported by statistical agencies for households, so it is difficult to differentiate between one-earner and two-earner income distributions. Some papers reported the use of interpolating functions with different asymptotic behavior for low and high incomes, such as the Tsallis function [116] and the Kaniadakis function [118]. However, the transition between the lower and upper classes is not smooth for the US data shown in Figs. 6 and 7, so such functions would not be useful in this case. The special case is income distribution in Argentina during the economic crisis, which shows a time-dependent bimodal shape with two peaks [116].

Theoretical Models of Income Distribution

Having examined the empirical data on income distribution, let us now discuss theoretical models. Income r_i is the influx of money per unit time to an agent i. If the money balance m_i is analogous to energy, then the income r_i would be analogous to power, which is the energy flux per unit time. So, one should conceptually distinguish between the distributions of money and income. While money is regularly transferred from one agent to another in pairwise transactions, it is not typical for agents to trade portions of their income. Nevertheless, indirect transfer of income may occur when one employee is promoted and another demoted, while the total annual budget is fixed, or when one company gets a contract, whereas another one loses it, etc. A reasonable approach, which has a long tradition in the economics literature [123,124,125], is to treat individual income r as a stochastic process and study its probability distribution. In general, one can study a Markov process generated by a matrix of transitions from one income to another. In the case where income r changes by a small amount Δr over a time period Δt, the Markov process can be treated as income diffusion. Then one can apply the general Fokker–Planck equation [71] to describe evolution in time t of the income distribution function $P(r,t)$ [105]:

$$\begin{aligned} &\frac{\partial P}{\partial t} = \frac{\partial}{\partial r}\left[AP + \frac{\partial(BP)}{\partial r}\right] , \\ &A = -\frac{\langle\Delta r\rangle}{\Delta t} , \quad B = \frac{\langle(\Delta r)^2\rangle}{2\Delta t} . \end{aligned} \tag{22}$$

The coefficients A and B in (22) are determined by the first and second moments of income changes per unit time. The stationary solution $\partial_t P = 0$ of (22) obeys the following equation with the general solution:

$$\begin{aligned} &\frac{\partial(BP)}{\partial r} = -AP , \\ &P(r) = \frac{c}{B(r)} \exp\left(-\int^r \frac{A(r')}{B(r')}\mathrm{d}r'\right) . \end{aligned} \tag{23}$$

For the lower part of the distribution, it is reasonable to assume that Δr is independent of r, i. e., the changes of income are independent of income itself. This process is called additive diffusion [105]. In this case, the coefficients in (22) are constants A_0 and B_0. Then (23) gives the exponential distribution $P(r) \propto \exp(-r/T_r)$, with the effective income temperature $T_r = B_0/A_0$. Notice that a meaningful stationary solution (23) requires that $A > 0$, i. e., $\langle\Delta r\rangle < 0$. The coincidence of this result with the Boltzmann–Gibbs exponential law (1) and (7) is not accidental. Indeed, instead of considering pairwise interaction between particles, one can derive (1) by considering energy transfers between a particle and a big reservoir, as long as the transfer process is "additive" and does not involve a Maxwell-demon-like discrimination. Stochastic income fluctuations are described by a similar process. So, although money and income are different concepts, they may have similar distributions, because they are governed by similar mathematical principles. It was shown explicitly in [25,82,83] that the models of pairwise money transfer can be described in a certain limit by the Fokker–Planck equation.

On the other hand, for the upper tail of the income distribution, it is reasonable to expect that $\Delta r \propto r$, i. e., income changes are proportional to income itself. This is known as the proportionality principle of Gibrat [123], and the process is called multiplicative diffusion [105]. In this case, $A = ar$ and $B = br^2$, and (23) gives the power-law distribution $P(r) \propto 1/r^{\alpha+1}$, with $\alpha = 1 + a/b$.

Generally, lower-class income comes from wages and salaries, where the additive process is appropriate, whereas upper-class income comes from bonuses, investments, and capital gains, calculated in percentages, where the multiplicative process applies [126]. However, the additive and multiplicative processes may coexist. An employee may receive a cost-of-living rise calculated in percentages (the multiplicative process) and a merit rise calculated in dollars (the additive process). In this case, we have $A = A_0 + ar$ and $B = B_0 + br^2 = b(r_0^2 + r^2)$, where $r_0^2 = B_0/b$. Substituting these expressions into (23), we find

$$P(r) = c\,\frac{\mathrm{e}^{-(\frac{r_0}{T_r})\arctan(\frac{r}{r_0})}}{[1+(\frac{r}{r_0})^2]^{\frac{1+a}{2b}}} . \tag{24}$$

The distribution (24) interpolates between the exponential law for low r and the power law for high r, because either the additive or the multiplicative process dominates in the corresponding limit. The crossover between the two regimes takes place at $r \sim r_0$, where the additive and multiplicative contributions to B are equal. The distribution (24) has three parameters: the "income temperature" $T_r = A_0/\mathrm{B}_0$, the Pareto exponent $\alpha = 1 + a/b$, and the crossover income r_0. It is a minimal model that captures the salient features of the empirical income distribution shown in Fig. 6. A mathematically similar, but more economically oriented model was proposed in [114,115], where labor income and asset accumulation are described by the additive and multiplicative processes correspondingly. A general stochastic process with additive and multiplicative noise was studied numerically in [127], but the stationary distribution was not derived analytically. A similar process with discrete time increments was studied by Kesten [128]. Recently, a formula similar to (24) was obtained in [129].

To verify the multiplicative and additive hypotheses empirically, it is necessary to have data on income mobility, i. e., the income changes Δr of the same people from one year to another. The distribution of income changes $P(\Delta r|r)$ conditional on income r is generally not available publicly, although it can be reconstructed by researchers at the tax agencies. Nevertheless, the multiplicative hypothesis for the upper class was quantitatively verified in [111,112] for Japan, where tax identification data is published for the top taxpayers.

The power-law distribution is meaningful only when it is limited to high enough incomes $r > r_0$. If all incomes r from 0 to ∞follow a purely multiplicative process, then one can change to a logarithmic variable $x = \ln(r/r_*)$ in (22) and show that it gives a Gaussian time-dependent distribution $P_t(x) \propto \exp(-x^2/2\sigma^2 t)$ for x, i. e., the log-normal distribution for r, also known as the Gibrat distribution [123]. However, the width of this distribution increases linearly in time, so the distribution is not stationary. This was pointed out by Kalecki [124] a long time ago, but the log-normal distribution is still widely used for fitting income distribution, despite this fundamental logical flaw in its justification. In a classic paper, Champernowne [125] showed that a multiplicative process gives a stationary power-law distribution when a boundary condition is imposed at $r_0 \neq 0$. Later, this result was rediscovered by econophysicists [130,131]. In our (24), the exponential distribution of the lower class effectively provides such a boundary condition for the power law of the upper class. Notice also that (24) reduces to (18) in the limit $r_0 \to 0$, which corresponds to purely multiplicative noise $B = br^2$.

There are alternative approaches to income distribution in the economics literature. One of them, proposed by Lydall [132], involves social hierarchy. Groups of people have leaders, who have leaders of a higher order, and so on. The number of people decreases geometrically (exponentially) with the increase of the hierarchical level. If individual income increases by a certain factor (i. e., multiplicatively) when moving to the next hierarchical level, then income distribution follows a power law [132]. However, the original argument of Lydall can be easily modified to produce an exponential distribution. If individual income increases by a certain amount, i. e., income increases linearly with the hierarchical level, then income distribution is exponential. The latter process seems to be more realistic for moderate incomes below \$ 100,000. A similar scenario is the Bernoulli trials [133], where individuals have a constant probability of increasing their income by a fixed amount. We see that the deterministic hierarchical models and the stochastic models of additive and multiplicative income mobility represent essentially the same ideas.

Other Applications of Statistical Physics

Statistical physics was applied to a number of other subjects in economics. Because of lack of space, only two such topics are briefly discussed in this section.

Economic Temperatures in Different Countries

As discussed in Sect. "Empirical Data on Money and Wealth Distributions"and "Empirical Data on Income Distribution", the distributions of money, wealth, and income are often described by exponential functions for the majority of the population. These exponential distributions are characterized by the parameters T_m, T_w, and T_r, which are mathematically analogous to temperature in the Boltzmann–Gibbs distribution (1). The values of these parameters, extracted from the fits of the empirical data, are generally different for different countries, i. e., different countries have different economic"temperatures". For example, Drăgulescu and Yakovenko [95] found that the US income temperature was 1.9 times higher than the UK income temperature in 1998 (using the exchange rate of dollars to pounds at that time). Also, there was ±25% variation between income temperatures of different states within the USA. [95].

In physics, a difference of temperatures allows one to set up a thermal machine. In was argued in [25] that the difference of money or income temperatures between different countries allows one to extract profit in international trade. Indeed, as discussed at the end of Sect. "Empirical Data on Money and Wealth Distributions", the prices of goods should be commensurate with money or income temperature, because otherwise people cannot afford to buy those goods. So, an international trading company can buy goods at a low price T_1 in a "low-temperature" country and sell them at a high price T_2 in a "high-temperature" country. The difference of prices $T_2 - T_1$ would be the profit of the trading company. In this process, money (the analog of energy) flows from the"high-temperature" to the "low-temperature" country, in agreement with the second law of thermodynamics, whereas products flow in the opposite direction. This process very much resembles what is going on in the global economy now. In this framework, the perpetual trade deficit of the USA is the consequence of the second law of thermodynamics and the difference of temperatures between the USA and "low-temperature"countries, such as China. Similar ideas were developed in more detail in [134,135], including a formal Carnot cycle for international trade.

The statistical physics approach demonstrates that profit originates from statistical nonequilibrium (the difference of temperatures), which exists in the global economy. However, it does not answer the question what is the origin of this difference. By analogy with physics, one would expect that the money flow should reduce the temperature difference and, eventually, lead to equilibrization of temperatures. In physics, this situation is known as the "thermal death of the universe". In a completely equilibrated global economy, it would be impossible to make profit by exploiting differences of economic temperatures between different countries. Although globalization of the modern economy does show a tendency toward equilibrization of living standards in different countries, this process is far from straightforward, and there are many examples contrary to equilibrization. This interesting and timely subject certainly requires further study.

Society as a Binary Alloy

In 1971, Thomas Schelling [136] proposed the now-famous mathematical model of segregation. He considered a lattice, where the sites can be occupied by agents of two types, e. g., blacks and whites in the problem of racial segregation. He showed that if the agents have some probabilistic preference for the neighbors of the same type, the system spontaneously segregates into black and white neighborhoods. This mathematical model is similar to the so-called Ising model, which is a popular model for studying phase transitions in physics. In this model, each lattice site is occupied by a magnetic atom, whose magnetic moment has only two possible orientations, up or down. The interaction energy between two neighboring atoms depends on whether their magnetic moments point in the same or in the opposite directions. In physics language, the segregation found by Schelling represents a phase transition in this system.

Another similar model is the binary alloy, a mixture of two elements which attract or repel each other. It was noticed in [137] that the behavior of actual binary alloys is strikingly similar to social segregation. In the following papers [42,138], this mathematical analogy was developed further and compared with social data. Interesting concepts, such as the coexistence curve between two phases and the solubility limit, were discussed in this work. The latter concept means that a small amount of one substance dissolves in another up to some limit, but phase separation (segregation) develops for higher concentrations. Recently, similar ideas were rediscovered in [139,140,141]. The vast experience of physicists in dealing with phase transitions and alloys may be helpful for practical applications of such models [142].

Future Directions, Criticism, and Conclusions

The statistical models described in this review are quite simple. It is commonly accepted in physics that theoretical models are not intended to be photographic copies of reality, but rather to be caricatures, capturing the most essential features of a phenomenon with a minimal number of details. With only few rules and parameters, the models discussed in Sect. "Statistical Mechanics of Money Distribution", "Statistical Mechanics of Wealth Distribution", and"Data and Models for Income Distribution" reproduce spontaneous development of stable inequality, which is present in virtually all societies. It is amazing that the calculated Gini coefficients, $G = 1/2$ for individuals and $G = 3/8$ for families, are actually very close to the US income data, as shown in Figs. 8 and 11. These simple models establish a baseline and a reference point for development of more sophisticated and more realistic models. Some of these future directions are outlined below.

Future Directions

Agents with a Finite Lifespan The models discussed in this review consider immortal agents who live forever, like atoms. However, humans have a finite lifespan. They enter the economy as young people and exit at an old age. Evolution of income and wealth as functions of age is studied in economics using the so-called overlapping-generations model. The absence of the age variable was one of the criticisms of econophysics by the economist Paul Anglin [31]. However, the drawback of the standard overlapping-generations model is that there is no variation of income and wealth between agents of the same age, because it is a representative-agent model. It would be best to combine stochastic models with the age variable. Also, to take into account inflation of average income, (22) should be rewritten for relative income, in the spirit of (17). These modifications would allow one to study the effects of demographic waves, such as baby boomers, on the distributions of income and wealth.

Agent-Based Simulations of the Two-Class Society The empirical data presented in Sect. "Empirical Data on Income Distribution" show quite convincingly that the US population consists of two very distinct classes characterized by different distribution functions. However, the theoretical models discussed in Sect. "Statistical Mechanics of Money Distribution"and "Statistical Mechanics of Wealth Distribution" do not produce two classes, although they do produce broad distributions. Generally, not much attention has been paid in the agent-based literature to simulation of two classes. One exception is [143], in which spontaneous development of employers and employees classes from initially equal agents was simulated [36]. More work in this direction would be certainly desirable.

Access to Detailed Empirical Data A great amount of statistical information is publicly available on the Internet, but not for all types of data. As discussed in Sect. "Empirical Data on Money and Wealth Distributions", it would be very interesting to obtain data on the distribution of balances on bank accounts, which would give information about the distribution of money (as opposed to wealth). As discussed in Sect. "Theoretical Models of Income Distribution", it would be useful to obtain detailed data on income mobility, to verify the additive and multiplicative hypotheses for income dynamics. Income distribution is often reported as a mix of data on individual income and family income, when the counting unit is a tax return (joint or single) or a household. To have a meaningful comparison with theoretical models, it is desirable to obtain clean data where the counting unit is an individual. Direct collaboration with statistical agencies would be very useful.

Economies in Transition Inequality in developed capitalist countries is generally quite stable. The situation is very different for former socialist countries making a transition to a market economy. According to the World Bank data [103], the average Gini coefficient for family income in eastern Europe and the former Soviet Union jumped from 25% in 1988 to 47% in 1993. The Gini coefficient in the socialist countries before the transition was well below the equilibrium value of 37.5% for market economies. However, the fast collapse of socialism left these countries out of market equilibrium and generated a much higher inequality. One may expect that, with time, their inequality will decrease to the equilibrium value of 37.5%. It would be very interesting to trace how fast this relaxation takes place. Such a study would also verify whether the equilibrium level of inequality is universal for all market economies.

Relation to Physical Energy The analogy between energy and money discussed in Sect. "Conservation of Money" is a formal mathematical analogy. However, actual physical energy with low entropy (typically in the form of fossil fuel) also plays a very important role in the modern economy, as the basis of current human technology. In view of the looming energy and climate crisis, it is imperative to find realistic ways for making a transition from the current "disposable" economy based on "cheap" and"unlimited" energy and natural resources to a sustainable one. Heterogeneity of human society is one of the important factors affecting such a transition. Econophysics, at the intersection of energy, entropy, economy, and statistical physics, may play a useful role in this quest [144].

Criticism from Economists

As econophysics is gaining popularity, some criticism has appeared from economists [31], including those who are closely involved with the econophysics movement [32,33,34]. This reflects a long-standing tradition in economic and social sciences of writing critiques on different schools of thought. Much of the criticism is useful and constructive and is already being accommodated in econophysics work. However, some criticism results from misunderstanding or miscommunication between the two fields and some from significant differences in scientific philosophy. Several insightful responses to the criticism have been published [145,146,147]; see also [7,148]. In this section, we briefly address the issues that are directly related to the material discussed in this review.

Awareness of Previous Economics Literature One complaint of [31,32,33,34] is that physicists are not well aware of the previous economics literature and either rediscover known results or ignore well-established approaches. To address this issue, it is useful to keep in mind that science itself is a complex system, and scientific progress is an evolutionary process with natural selection. The sea of scientific literature is enormous, and nobody knows it all. Recurrent rediscovery of regularities in the natural and social world only confirms their validity. Independent rediscovery usually brings a different perspective, broader applicability range, higher accuracy, and better mathematical treatment, so there is progress even when some overlap with previous results exists. Physicists are grateful to economists for bringing relevant and specific references to their attention. Since the beginning of modern econophysics, many old references have been uncovered and are now routinely cited.

However, not all old references are relevant to the new development. For example, Gallegati et al. [33] complained that the econophysics literature on income distribution ignores the so-called Kuznets hypothesis [149]. The Kuznets hypothesis postulates that income inequality first rises during an industrial revolution and then decreases, producing an inverted-U-shaped curve. Gallegati et al. [33] admitted that, to date, the large amount of literature on the Kuznets hypothesis is inconclusive. Kuznets [149] mentioned that this hypothesis applies to the period from colonial times to the 1970s; however, the empirical data for this period are sparse and not very reliable. The econophysics literature deals with reliable volumes of data for the second half of the twentieth century, collected with the introduction of computers. It is not clear what is the definition of industrial revolution and when exactly it starts and ends. The chain of technological progress seems to be continuous (steam engine, internal combustion engine, cars, plastics, computers, Internet), so it is not clear where the purported U-curve is supposed to be placed in time. Thus, the Kuznets hypothesis appears to be, in principle, unverifiable and unfalsifiable. The original paper by Kuznets [149] actually does not contain any curves, but it has one table filled with made-up, imaginary data! Kuznets admits that he has "neither the necessary data nor a reasonably complete theoretical model" (p. 12 in [149]). So, this paper is understandably ignored by the econophysics community. In fact, the data analysis for 1947–1984 shows amazing stability of income distribution [150], consistent with Fig. 11. The increase of inequality in the 1990s resulted from growth of the upper class relative to the lower class, but the relative inequality within the lower class remains very stable, as shown in Fig. 7.

Reliance on Visual Data Analysis Another complaint of [33] is that econophysicists favor graphic analysis of data over the formal and "rigorous" testing prescribed by mathematical statistics, as favored by economists. This complaint goes against the trend of all sciences to use increasingly sophisticated data visualization for uncovering regularities in complex systems. The thick IRS publication 1304 [151] is filled with data tables, but has virtually no graphs. Despite the abundance of data, it gives a reader no idea about income distribution, whereas plotting the data immediately gives insight. However, intelligent plotting is the art with many tools, which not many researchers have mastered. The author completely agrees with Gallegati et al. [33] that too many papers mindlessly plot any kind of data on a log–log scale, pick a finite interval, where any smooth curved line can be approximated by a straight line, and claim that there is a power law. In many cases, replotting the same data on a log–linear scale converts a curved line into a straight line, which means that the law is actually exponential.

Good visualization is extremely helpful in identifying trends in complex data, which can then be fitted to a mathematical function; however, for a complex system, such a fit should not be expected with infinite precision. The fundamental laws of physics, such as Newton's law of gravity or Maxwell's equations, are valid with enormous precision. However, the laws in condensed matter physics, uncovered by experimentalists with a combination of visual analysis and fitting, usually have much lower precision, at best 10% or so. Most of these laws would fail the formal criteria of mathematical statistics. Nevertheless these approximate laws are enormously useful in practice, and everyday devices engineered on the basis of these laws work very well for all of us.

Because of the finite accuracy, different functions may produce equally good fits. Discrimination between the exponential, Gamma, and log-normal functions may not be always possible [122]. However, the exponential function has fewer fitting parameters, so it is preferable on the basis of simplicity. The other two functions can simply mimic the exponential function with a particular choice of the additional parameters [122]. Unfortunately, many papers in mathematical statistics introduce too many fitting parameters into complicated functions, such as the generalized beta distribution mentioned in [33]. Such overparameterization is more misleading than insightful for data fitting.

Quest for Universality Gallegati et al. [33] criticized physicists for trying to find universality in economics data. They also seemed to equate the concepts of power law, scaling, and universality. These are three different, albeit overlapping, concepts. Power laws usually apply only to a small fraction of data at the high ends of various distributions. Moreover, the exponents of these power laws are usually nonuniversal and vary from case to case. Scaling means that the shape of a function remains the same when its scale changes. However, the scaling function does not have to be a power-law function. A good example of scaling is shown in Fig. 7, where income distributions for the lower class collapse on the same exponential line for about 20 years of data. We observe amazing universality of income distribution, unrelated to a power law. In a general sense, the diffusion equation is universal, because it describes a wide range of systems, from dissolution of sugar in water to a random walk in the stock market.

Universalities are not easy to uncover, but they form the backbone of regularities in the world around us. This is why physicists are so interested in them. Universalities establish the first-order effect, and deviations represent the second-order effect. Different countries may have somewhat different distributions, and economists often tend to focus on these differences. However, this focus on details misses the big picture that, in the first approximation, the distributions are quite similar and universal.

Theoretical Modeling of Money, Wealth, and Income
It was pointed out in [31,33,34] that many econophysics papers confuse or misuse the terms for money, wealth, and income. It is true that terminology is sloppy in many papers and should be refined. However, the terms in [25,26] are quite precise, and this review clearly distinguishes between these concepts in Sect. "Statistical Mechanics of Money Distribution", "Statistical Mechanics of Wealth Distribution", and "Data and Models for Income Distribution".

One contentious issue is about conservation of money. Gallegati et al. [33] agree that"transactions are a key economic process, and they are necessarily conservative", i. e., money is indeed conserved in transactions between agents. However, Anglin [31], Gallegati et al. [33], and Lux [34] complain that the models of conservative exchange do not consider production of goods, which is the core economic process and the source of economic growth. Material production is indeed the ultimate goal of an economy, but it does not violate conservation of money by itself. One can grow coffee beans, but nobody can grow money on a money tree. Money is an artificial economic device that is designed to be conserved. As explained in Sect. "Statistical Mechanics of Money Distribution", the money transfer models implicitly assume that money in transactions is voluntarily paid for goods and services generated by production for the mutual benefit of the parties. In principle, one can introduce a billion variables to keep track of every coffee bean and other product of the economy. What difference would it make for the distribution of *money*? Despite the claims in [31,33], there is no contradiction between models of conservative exchange and the classic work of Adam Smith and David Ricardo. The difference is only in the focus: We keep track of money, whereas they keep track of coffee beans, from production to consumption. These approaches address different questions, but do not contradict each other. Because money constantly circulates in the system as payment for production and consumption, the resulting statistical distribution of money may very well not depend on what exactly is produced and in what quantities.

In principle, the models with random transfers of money should be considered as a reference point for developing more sophisticated models. Despite the totally random rules and "zero intelligence" of the agents, these models develop well-characterized, stable, and stationary distributions of money. One can modify the rules to make the agents more intelligent and realistic and see how much the resulting distribution changes relative to the reference one. Such an attempt was made in [32] by modifying the model of [77] with various more realistic economic ingredients. However, despite the modifications, the resulting distributions were essentially the same as in the original model. This example illustrates the typical robustness and universality of statistical models: Modifying details of microscopic rules does not necessarily change the statistical outcome.

Another misconception, elaborated in [32,34], is that the money transfer models discussed in Sect. "Statistical Mechanics of Money Distribution" imply that money is transferred by fraud, theft, and violence, rather than voluntarily. One should keep in mind that the catchy labels "theft-and-fraud", "marriage-and-divorce", and "yard-sale" were given to the money transfer models by the journalist Brian Hayes [152] in a popular article. Econophysicists who originally introduced and studied these models do not subscribe to this terminology, although the early work of Angle [51] did mention violence as one source of redistribution. In the opinion of the author, it is indeed difficult to justify the proportionality rule (8), which implies that agents with high balances pay proportionally greater amounts in transactions than agents with low balances. However, the additive model of [25], where money transfers Δm are independent of money balances m_i of the agents, does not have this problem. As explained in Sect. "The Boltzmann–Gibbs Distribution of Money", this model simply means that all agents pay the same prices for the same product, although prices may be different for different products. So, this model is consistent with voluntary transactions in a free market.

McCauley [145] argued that conservation of money is violated by credit. As explained in Sect. "Models with Debt", credit does not violate conservation law, but creates positive and negative money without changing net worth. Negative money (debt) is as real as positive money. McCauley [145] claimed that money can be easily created with the tap of a computer key via credit. Then why would an employer not tap the key and double salaries, or a funding agency double research grants? Because budget constraints are real. Credit may provide a temporary relief, but sooner or later it has to be paid back. Allowing debt may produce a double-exponential distribution as shown in Fig. 3, but it doesn't change the distribution fundamentally.

As discussed in Sect. "Conservation of Money", a central bank or a central government can inject new money into the economy. As discussed in Sect. "Statistical Mechanics of Wealth Distribution", wealth is generally not conserved. As discussed in Sect. "Data and Models for Income Distribution", income is different from money and is described by a different model (22). However, the empirical distribution of income shown in Fig. 6 is qualitatively similar to the distribution of wealth shown in Fig. 5, and we do not have data on money distribution.

Conclusions

The "invasion" of physicists into economics and finance at the turn of the millennium is a fascinating phenomenon. The physicist Joseph McCauley proclaims that "Econophysics will displace economics in both the universities and boardrooms, simply because what is taught in economics classes doesn't work" [153]. Although there is some truth in his arguments [145], one may consider a less radical scenario. Econophysics may become a branch of economics, in the same way as games theory, psychological economics, and now agent-based modeling became branches of economics. These branches have their own interests, methods, philosophy, and journals. The main contribution from the infusion of new ideas from a different field is not in answering old questions, but in raising new questions. Much of the misunderstanding between economists and physicists happens not because they are getting different answers, but because they are answering different questions.

The subject of income and wealth distributions and social inequality was very popular at the turn of another century and is associated with the names of Pareto, Lorenz, Gini, Gibrat, and Champernowne, among others. Following the work by Pareto, attention of researchers was primarily focused on the power laws. However, when physicists took a fresh, unbiased look at the empirical data, they found a different, exponential law for the lower part of the distribution. The motivation for looking at the exponential law, of course, came from the Boltzmann–Gibbs distribution in physics. Further studies provided a more detailed picture of the two-class distribution in a society. Although social classes have been known in political economy since Karl Marx, the realization that they are described by simple mathematical distributions is quite new. Demonstration of the ubiquitous nature of the exponential distribution for money, wealth, and income is one of the new contributions produced by econophysics.

Bibliography

Primary Literature

1. Chakrabarti BK (2005) Econophys-Kolkata: a short story. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of Wealth Distributions. Springer, Milan, pp 225–228
2. Carbone A, Kaniadakis G, Scarfone AM (2007) Where do we stand on econophysics? Phys A 382:xi–xiv
3. Stanley HE et al (1996) Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. Phys A 224:302–321
4. Mantegna RN, Stanley HE (1999) An introduction to econophysics: correlations and complexity in finance. Cambridge University Press, Cambridge
5. Galam S (2004) Sociophysics: a personal testimony. Phys A 336:49–55
6. Galan S, Gefen Y, Shapir Y (1982) Sociophysics: a new approach of sociological collective behaviour. I. Mean-behaviour description of a strike. J Math Soc 9:1–13
7. Stauffer D (2004) Introduction to statistical physics outside physics. Phys A 336:1–5
8. Schweitzer F (2003) Brownian agents and active particles: collective dynamics in the natural and social sciences. Springer, Berlin
9. Weidlich W (2000) Sociodynamics: a systematic approach to mathematical modeling in the social sciences. Harwood Academic Publishers, Amsterdam
10. Chakrabarti BK, Chakraborti A, Chatterjee A (eds) (2006) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
11. Ball P (2002) The physical modelling of society: a historical perspective. Phys A 314:1–14
12. Ball P (2004) Critical mass: how one thing leads to another. Farrar, Straus and Giroux, New York
13. Boltzmann L (1905) Populäre Schriften. Barth, Leipzig, p 360
14. Austrian Central Library for Physics (2006) Ludwig Boltzmann 1844–1906. ISBN 3-900490-11-2. Vienna
15. Pareto V (1897) Cours d'Économie Politique. L'Université de Lausanne
16. Mirowski P (1989) More heat than light: economics as social physics, physics as nature's economics. Cambridge University Press, Cambridge
17. Majorana E (1942) Il valore delle leggi statistiche nella fisica e nelle scienze sociali. Scientia 36:58–66 (English translation by Mantegna RN in: Bassani GF (ed) (2006) Ettore Majorana Scientific Papers. Springer, Berlin, pp 250–260)
18. Montroll EW, Badger WW (1974) Introduction to quantitative aspects of social phenomena. Gordon and Breach, New York
19. Föllmer H (1974) Random economies with many interacting agents. J Math Econ 1:51–62
20. Blume LE (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5:387–424
21. Foley DK (1994) A statistical equilibrium theory of markets. J Econ Theory 62:321–345
22. Durlauf SN (1997) Statistical mechanics approaches to socioeconomic behavior. In: Arthur WB, Durlauf SN, Lane DA (eds) The Economy as a Complex Evolving System II. Addison-Wesley, Redwood City, pp 81–104
23. Anderson PW, Arrow KJ, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Reading
24. Rosser JB (2008) Econophysics. In: Blume LE, Durlauf SN (eds) New Palgrave Dictionary of Economics, 2nd edn. Macmillan, London (in press)
25. Drăgulescu AA, Yakovenko VM (2000) Statistical mechanics of money. Europ Phys J B 17:723–729

26. Chakraborti A, Chakrabarti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. Europ Phys J B 17:167–170
27. Bouchaud JP, Mézard M (2000) Wealth condensation in a simple model of economy. Phys A 282:536–545
28. Wannier GH (1987) Statistical physics. Dover, New York
29. Lopez-Ruiz R, Sanudo J, Calbet X (2007) Geometrical derivation of the Boltzmann factor. Available via DIALOG. http://arxiv.org/abs/0707.4081. Accessed 1 Jul 2008
30. Lopez-Ruiz R, Sanudo J, Calbet X (2007) On the equivalence of the microcanonical and the canonical ensembles: a geometrical approach. Available via DIALOG. http://arxiv.org/abs/0708.1866. Accessed 1 Jul 2008
31. Anglin P (2005) Econophysics of wealth distributions: a comment. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, New York, pp 229–238
32. Lux T (2005) Emergent statistical wealth distributions in simple monetary exchange models: a critical review. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 51–60
33. Gallegati M, Keen S, Lux T, Ormerod P (2006) Worrying trends in econophysics. Phys A 370:1–6
34. Lux T (2008) Applications of statistical physics in finance and economics. In: Rosser JB (ed) Handbook of complexity research. Edward Elgar, Cheltenham, UK and Northampton, MA (in press)
35. Computer animation videos of money-transfer models. http://www2.physics.umd.edu/~yakovenk/econophysics/animation.html. Accessed 1 Jul 2008
36. Wright I (2007) Computer simulations of statistical mechanics of money in mathematica. Available via DIALOG. http://demonstrations.wolfram.com/StatisticalMechanicsOfMoney. Accessed 1 Jul 2008
37. McConnell CR, Brue SL (1996) Economics: principles, problems, and policies. McGraw-Hill, New York
38. Patriarca M, Chakraborti A, Kaski K, Germano G (2005) Kinetic theory models for the distribution of wealth: Power law from overlap of exponentials. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 93–110
39. Bennati E (1988) Un metodo di simulazione statistica per l'analisi della distribuzione del reddito. Rivista Internazionale di Scienze Economiche e Commerciali 35:735–756
40. Bennati E (1993) Il metodo di Montecarlo nell'analisi economica. Rassegna di Lavori dell'ISCO (Istituto Nazionale per lo Studio della Congiuntura), Anno X 4:31–79
41. Scalas E, Garibaldi U, Donadio S (2006) Statistical equilibrium in simple exchange games I: methods of solution and application to the Bennati–Drăgulescu–Yakovenko (BDY) game. Europ Phys J B 53:267–272
42. Mimkes J (2000) Society as a many-particle system. J Therm Anal Calorim 60:1055–1069
43. Mimkes J (2005) Lagrange principle of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 61–69
44. Shubik M (1999) The theory of money and financial institutions, vol 2. MIT Press, Cambridge, p 192
45. Mandelbrot B (1960) The Pareto-Lévy law and the distribution of income. Int Econ Rev 1:79–106
46. Braun D (2001) Assets and liabilities are the momentum of particles and antiparticles displayed in Feynman-graphs. Phys A 290:491–500
47. Fischer R, Braun D (2003) Transfer potentials shape and equilibrate monetary systems. Phys A 321:605–618
48. Fischer R, Braun D (2003) Nontrivial bookkeeping: a mechanical perspective. Phys A 324:266–271
49. Xi N, Ding N, Wang Y (2005) How required reserve ratio affects distribution and velocity of money. Phys A 357:543–555
50. Ispolatov S, Krapivsky PL, Redner S (1998) Wealth distributions in asset exchange models. Europ Phys J B 2:267–276
51. Angle J (1986) The surplus theory of social stratification and the size distribution of personal wealth. Soc Forces 65:293–326
52. Angle J (1992) The inequality process and the distribution of income to blacks and whites. J Math Soc 17:77–98
53. Angle J (1992) Deriving the size distribution of personal wealth from 'the rich get richer, the poor get poorer'. J Math Soc 18:27–46
54. Angle J (1996) How the Gamma Law of income distribution appears invariant under aggregation. J Math Soc 21:325–358
55. Angle J (2002) The statistical signature of pervasive competition on wage and salary incomes. J Math Soc 26:217–270
56. Angle J (2006) The Inequality Process as a wealth maximizing process. Phys A 367:388–414
57. Engels F (1972) The origin of the family, private property and the state, in the light of the researches of Lewis H. Morgan. International Publishers, New York
58. Patriarca M, Chakraborti A, Kaski K (2004) Gibbs versus non-Gibbs distributions in money dynamics. Phys A 340:334–339
59. Patriarca M, Chakraborti A, Kaski K (2004) Statistical model with a standard Gamma distribution. Phys Rev E 70:016104
60. Repetowicz P, Hutzler S, Richmond P (2005) Dynamics of money and income distributions. Phys A 356:641–654
61. Chatterjee A, Chakrabarti BK, Manna SS (2004) Pareto law in a kinetic model of market with random saving propensity. Phys A 335:155-163
62. Das A, Yarlagadda S (2005) An analytic treatment of the Gibbs-Pareto behavior in wealth distribution. Phys A 353:529–538
63. Chatterjee S, Chakrabarti BK, Stinchcombe RB (2005) Master equation for a kinetic model of a trading market and its analytic solution. Phys Rev E 72:026126
64. Mohanty PK (2006) Generic features of the wealth distribution in ideal-gas-like markets. Phys Rev E 74:011117
65. Patriarca M, Chakraborti A, Germano G (2006) Influence of saving propensity on the power-law tail of the wealth distribution. Phys A 369:723–736
66. Gupta AK (2006) Money exchange model and a general outlook. Phys A 359:634–640
67. Patriarca M, Chakraborti A, Heinsalu E, Germano G (2007) Relaxation in statistical many-agent economy models. Europ Phys J B 57:219–224
68. Ferrero JC (2004) The statistical distribution of money and the rate of money transference. Phys A 341:575–585
69. Scafetta N, Picozzi S, West BJ (2004) An out-of-equilibrium model of the distributions of wealth. Quant Financ 4:353–364
70. Scafetta N, Picozzi S, West BJ (2004) A trade-investment model for distribution of wealth. Physica D 193:338–352
71. Lifshitz EM, Pitaevskii LP (1981) Physical kinetics. Pergamon Press, Oxford

72. Ao P (2007) Boltzmann–Gibbs distribution of fortune and broken time reversible symmetry in econodynamics. Commun Nonlinear Sci Numer Simul 12:619–626
73. Scafetta N, West BJ (2007) Probability distributions in conservative energy exchange models of multiple interacting agents. J Phys Condens Matter 19:065138
74. Chakraborti A, Pradhan S, Chakrabarti BK (2001) A self-organising model of market with single commodity. Phys A 297:253–259
75. Chatterjee A, Chakrabarti BK (2006) Kinetic market models with single commodity having price fluctuations. Europ Phys J B 54:399–404
76. Ausloos M, Pekalski A (2007) Model of wealth and goods dynamics in a closed market. Phys A 373:560–568
77. Silver J, Slud E, Takamoto K (2002) Statistical equilibrium wealth distributions in an exchange economy with stochastic preferences. J Econ Theory 106:417–435
78. Raberto M, Cincotti S, Focardi SM, Marchesi M (2003) Traders' long-run wealth in an artificial financial market. Comput Econ 22:255–272
79. Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. Phys A 299:188–197
80. Solomon S, Richmond P (2002) Stable power laws in variable economies; Lotka-Volterra implies Pareto-Zipf. Europ Phys J B 27:257–261
81. Huang DW (2004) Wealth accumulation with random redistribution. Phys Rev E 69:057103
82. Slanina F (2004) Inelastically scattering particles and wealth distribution in an open economy. Phys Rev E 69:046102
83. Cordier S, Pareschi L, Toscani G (2005) On a kinetic model for a simple market economy. J Statist Phys 120:253–277
84. Richmond P, Repetowicz P, Hutzler S, Coelho R (2006) Comments on recent studies of the dynamics and distribution of money. Phys A 370:43–48
85. Richmond P, Hutzler S, Coelho R, Repetowicz P (2006) A review of empirical studies and models of income distributions in society. In: Chakrabarti BK, Chakraborti A Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
86. Burda Z, Johnston D, Jurkiewicz J, Kaminski M, Nowak MA, Papp G, Zahed I (2002) Wealth condensation in Pareto macroeconomies. Phys Rev E 65:026102
87. Pianegonda S, Iglesias JR, Abramson G, Vega JL (2003) Wealth redistribution with conservative exchanges. Phys A 322:667–675
88. Braun D (2006) Nonequilibrium thermodynamics of wealth condensation. Phys A 369:714–722
89. Coelho R, Néda Z, Ramasco JJ, Santos MA (2005) A family-network model for wealth distribution in societies. Phys A 353:515–528
90. Iglesias JR, Gonçalves S, Pianegonda S, Vega JL, Abramson G (2003) Wealth redistribution in our small world. Phys A 327:12–17
91. Di Matteo T, Aste T, Hyde ST (2004) Exchanges in complex networks: income and wealth distributions. In: Mallamace F, Stanley HE (eds) The physics of complex systems (New advances and perspectives). IOS Press, Amsterdam, p 435
92. Hu MB, Jiang R, Wu QS, Wu YH (2007) Simulating the wealth distribution with a Richest-Following strategy on scale-free network. Phys A 381:467–472
93. Bak P, Nørrelykke SF, Shubik M (1999) Dynamics of money. Phys Rev E 60:2528–2532
94. Her Majesty Revenue and Customs (2003) Distribution of personal wealth. Available via DIALOG. http://www.hmrc.gov.uk/stats/personal_wealth/wealth_oct03.pdf. Accessed 1 Jul 2008
95. Drăgulescu AA, Yakovenko VM (2001) Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Phys A 299:213–221
96. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2007) The Forbes 400, the Pareto power-law and efficient markets. Europ Phys J B 55:143–147
97. Sinha S (2006) Evidence for power-law tail of the wealth distribution in India. Phys A 359:555–562
98. Abul-Magd AY (2002) Wealth distribution in an ancient Egyptian society. Phys Rev E 66:057104
99. Kakwani N (1980) Income Inequality and Poverty. Oxford University Press, Oxford
100. Champernowne DG, Cowell FA (1998) Economic inequality and income distribution. Cambridge University Press, Cambridge
101. Atkinson AB, Bourguignon F (eds) (2000) Handbook of income distribution. Elsevier, Amsterdam
102. Drăgulescu AA, Yakovenko VM (2001) Evidence for the exponential distribution of income in the USA. Europ Phys J B 20:585–589
103. Drăgulescu AA, Yakovenko VM (2003) Statistical mechanics of money, income, and wealth: a short survey. In: Garrido PL, Marro J (eds) Modeling of complex systems: seventh granada lectures, Conference Proceedings 661. American Institute of Physics, New York, pp 180–183
104. Yakovenko VM, Silva AC (2005) Two-class structure of income distribution in the USA: Exponential bulk and power-law tail. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 15–23
105. Silva AC, Yakovenko VM (2005) Temporal evolution of the 'thermal' and 'superthermal' income classes in the USA during 1983-2001. Europhys Lett 69:304–310
106. Hasegawa A, Mima K, Duong-van M (1985) Plasma distribution function in a superthermal radiation field. Phys Rev Lett 54:2608–2610
107. Desai MI, Mason GM, Dwyer JR, Mazur JE, Gold RE, Krimigis SM, Smith CW, Skoug RM (2003) Evidence for a suprathermal seed population of heavy ions accelerated by interplanetary shocks near 1 AU. Astrophys J 588:1149–1162
108. Collier MR (2004) Are magnetospheric suprathermal particle distributions (κ functions) inconsistent with maximum entropy considerations? Adv Space Res 33:2108–2112
109. Souma W (2001) Universal structure of the personal income distribution. Fractals 9:463–470
110. Souma W (2002) Physics of personal income. In: Takayasu H (ed) Empirical science of financial fluctuations: the advent of econophysics. Springer, Tokyo, pp 343–352
111. Fujiwara Y, Souma W, Aoyama H, Kaizoji T, Aoki M (2003) Growth and fluctuations of personal income. Phys A 321:598–604
112. Aoyama H, Souma W, Fujiwara Y (2003) Growth and fluctuations of personal and company's income. Phys A 324:352–358
113. Strudler M, Petska T, Petska R (2003) An analysis of the distribution of individual income and taxes, 1979–2001. The In-

ternal Revenue Service, Washington DC. Available via DIALOG. http://www.irs.gov/pub/irs-soi/03strudl.pdf. Accessed 1 Jul 2008

114. Souma W, Nirei M (2005) Empirical study and model of personal income. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 34–42
115. Nirei M, Souma W (2007) A two factor model of income distribution dynamics. Rev Income Wealth 53:440–459
116. Ferrero JC (2005) The monomodal, polymodal, equilibrium and nonequilibrium distribution of money. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 159–167
117. Clementi F, Gallegati M (2005) Pareto's law of income distribution: evidence for Germany, the United Kingdom, the United States. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 3–14
118. Clementi F, Gallegati M, Kaniadakis G (2007) κ-generalized statistics in personal income distribution. Europ Phys J B 57:187–193
119. Clementi F, Gallegati M (2005) Power law tails in the Italian personal income distribution. Phys A 350:427–438
120. Clementi F, Di Matteo T, Gallegati M (2006) The power-law tail exponent of income distributions. Phys A 370:49–53
121. Rawlings PK, Reguera D, Reiss H (2004) Entropic basis of the Pareto law. Phys A 343:643–652
122. Banerjee A, Yakovenko VM, Di Matteo T (2006) A study of the personal income distribution in Australia. Phys A 370:54–59
123. Gibrat R (1931) Les Inégalités Economiques. Sirely, Paris
124. Kalecki M (1945) On the Gibrat distribution. Econometrica 13:161–170
125. Champernowne DG (1953) A model of income distribution. Econ J 63:318–351
126. Milaković M (2005) Do we all face the same constraints? In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 184–191
127. Takayasu H, Sato AH, Takayasu M (1997) Stable infinite variance fluctuations in randomly amplified Langevin systems. Phys Rev Lett 79:966–969
128. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Math 131:207–248
129. Fiaschi D, Marsili M (2007) Distribution of wealth: theoretical microfoundations and empirical evidence. Working paper. Avialable via DIALOG. http://www.dse.ec.unipi.it/persone/docenti/fiaschi/Lavori/distributionWealthMicrofoundations.pdf. Accessed 1 Jul 2008
130. Levy M, Solomon S (1996) Power laws are logarithmic Boltzmann laws. Int J Mod Phys C 7:595–751
131. Sornette D, Cont R (1997) Convergent multiplicative processes repelled from zero: power laws and truncated power laws. J Phys I (France) 7:431–444
132. Lydall HF (1959) The distribution of employment incomes. Econometrica 27:110–115
133. Feller W (1966) An Introduction to Probability Theory and Its Applications, vol 2. Wiley, New York, p 10
134. Mimkes J, Aruka Y (2005) Carnot process of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 70–78
135. Mimkes J (2006) A thermodynamic formulation of economics. In: Chakrabarti BK, Chakraborti A Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin, pp 1–33
136. Schelling TC (1971) Dynamic models of segregation. J Math Soc 1:143–186
137. Mimkes J (1995) Binary alloys as a model for the multicultural society. J Therm Anal 43:521–537
138. Mimkes J (2006) A thermodynamic formulation of social science. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
139. Jego C, Roehner BM (2007) A physicist's view of the notion of "racism". Available via DIALOG. http://arxiv.org/abs/0704.2883. Accessed 1 Jul 2008
140. Stauffer D, Schulze C (2007) Urban and scientific segregation: the Schelling-Ising model. Avialable via DIALOG. http://arxiv.org/abs/0710.5237. Accessed 1 Jul 2008
141. Dall'Asta L, Castellano C, Marsili M (2007) Statistical physics of the Schelling model of segregation. Available via DIALOG. http://arxiv.org/abs/0707.1681. Accessed 1 Jul 2008
142. Lim M, Metzler R, Bar-Yam Y (2007) Global pattern formation and ethnic/cultural violence. Science 317:1540–1544
143. Wright I (2005) The social architecture of capitalism. Phys A 346:589–620
144. Defilla S (2007) A natural value unit – Econophysics as arbiter between finance and economics. Phys A 382:42–51
145. McCauley JL (2006) Response to 'Worrying Trends in Econophysics'. Phys A 371:601–609
146. Richmond P, Chakrabarti BK, Chatterjee A, Angle J (2006) Comments on 'Worrying Trends in Econophysics': income distribution models. In: Chatterjee A, Chakrabarti BK (eds) Econophysics of stock and other markets. Springer, Milan, pp 244–253
147. Rosser JB (2006) Debating the Role of Econophysics. Working paper. Available via DIALOG. http://cob.jmu.edu/rosserjb/. Accessed 1 Jul 2008
148. Rosser JB (2006) The nature and future of econophysics. In: Chatterjee A, Chakrabarti BK (eds) Econophysics of stock and other markets. Springer, Milan, pp 225–234
149. Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45:1–28
150. Levy F (1987) Changes in the distribution of American family incomes, 1947 to 1984. Science 236:923–927
151. Internal Revenue Service (1999) Statistics of Income–1997, Individual Income Tax Returns. Publication 1304, Revision 12-99, Washington DC
152. Hayes B (2002) Follow the money. Am Sci 90:400–405
153. Ball P (2006) Econophysics: culture crash. Nature 441:686–688

Books and Reviews

McCauley J (2004) Dynamics of markets: econophysics and finance. Cambridge University Press, Cambridge

Farmer JD, Shubik M, Smith E (2005) Is economics the next physical science? Phys Today 58(9):37–42

Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Agent-based models of financial markets. Rep Prog Phys 70:409–450

Econophysics forum. Avialable via DIALOG. http://www.unifr.ch/econophysics/. Accessed 1 Jul 2008

Extreme Events in Socio-economic and Political Complex Systems, Predictability of

Vladimir Keilis-Borok[1,2], Alexandre Soloviev[2,3], Allan Lichtman[4]

[1] Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA

[2] International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia

[3] Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

[4] American University, Washington D.C., USA

Article Outline

Glossary

Complexity A definitive feature of nonlinear systems of interacting elements. It comprises high instability with respect to initial and boundary conditions, and complex but non-random behavior patterns ("order in chaos").

Extreme events Rare events having a large impact. Such events are also known as critical phenomena, disasters, catastrophes, and crises. They persistently reoccur in hierarchical complex systems created, separately or jointly, by nature and society.

Fast acceleration of unemployment (FAU) The start of a strong and lasting increase of the unemployment rate.

Pattern recognition of rare events The methodology of artificial intelligence' kind aimed at studying distinctive features of complex phenomena, in particular – at formulating and testing hypotheses on these features.

Premonitory patterns Patterns of a complex system's behavior that emerge most frequently as an extreme event approaches.

Recession The American National Bureau of Economic Research defines recession as "a significant decline in economic activity spread across the economy, lasting more than a few months". A recession may involve simultaneous decline in coincident measures of overall economic activity such as industrial production, employment, investment, and corporate profits.

Start of the homicide surge (SHS) The start of a strong and lasting increase in the smoothed homicide rate.

Definition of the Subject

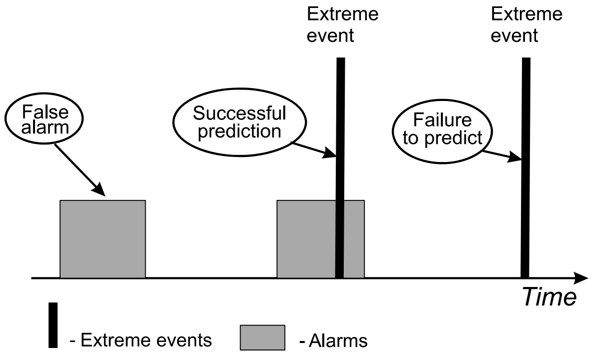
At stake in the development of accurate and reliable methods of prediction for social systems is the capacity of scientific reason to improve the human condition. Today's civilization is highly vulnerable to crises arising from extreme events generated by complex and poorly understood systems. Examples include external and civil wars, terrorist attacks, crime waves, economic downturns, and famines, to name just a few. Yet more subtle effects threaten modern society, such as the inability of democratic systems to produce policies responsive to challenges like climate change, global poverty, and resource depletion.

Our capacity to predict the course of events in complex social systems is inherently limited. However, there is a new and promising approach to predicting and understanding complex systems that has emerged through the integration of studies in the social sciences and the mathematics of prediction. This entry describes and analyzes that approach and its real-world applications. These include algorithmic prediction of electoral fortunes of incumbent parties, economic recessions, surges of unemployment, and outbursts of crimes. This leads to important inferences for averting and responding to impending crises and for improving the functioning of modern democratic societies.

That approach was successfully applied also to natural disasters such as earthquakes. Ultimately, improved prediction methods enhance our capacity for understanding the world and for protecting and sustaining our civilization.

Extreme events. Hierarchical complex systems persistently generate extreme events – the rare fast changes that have a strong impact on the system. Depending on connotation they are also known as critical phenomena, disasters, catastrophes, and crises. This article examines the development and application of the algorithmic prediction of extreme socio-economic and political events.

The prediction problem is formulated as follows:
given are time series that describe dynamics of the sys-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 1
Possible outcomes of prediction

tem up to the current moment of time t and contain potential precursors of an extreme event;

to predict whether an extreme event will or will not occur during the subsequent time period $(t, t + \tau)$; if the answer is "yes", this will be the "*period of alarm*".

As the time goes by, predictions form a discrete sequence of alarms. The possible outcomes of such a prediction are shown in Fig. 1. The actual outcome is determined unambiguously, since the extreme events are identified independently of the prediction either by the actual happening (e. g. by an election result) or by a separate algorithm (e. g. homicide surge) after they occur.

Such "yes or no" prediction is aimed not at analyzing the whole dynamics of the system, but only at identifying the occurrence of rare extreme events. In a broad field of prediction studies this prediction is different from and complementary to the classical Kolmogoroff–Wiener prediction of continuous functions, and to traditional cause-and-effect analysis.

The problem includes estimating the predictions' accuracy: the rates of false alarms and failures to predict, and the total duration of alarms in relation to the total time considered. These characteristics represent the inevitable *probabilistic component* of prediction; they provide for statistical validation of a prediction algorithm and for optimizing preparedness to predicted events (e. g. recessions or crime surges).

Twofold importance. The prediction problem is pivotal in two areas:

- *Fundamental understanding of complex systems*. Prediction algorithms quantitatively define phenomena that anticipate extreme events. Such quantitative definition is pivotal for fundamental understanding of a complex system where these events occur, including the intertwined mechanisms of system's development and its basic features, e. g. multiple scaling, correlation range, clustering, fragmentation etc. (see Sects. "Common Elements of Data Analyzes", "Elections", "US Economic Recessions", "Unemployment"). The understanding of complex systems remains a major unsolved problem of modern science, tantamount to transforming our understanding of the natural and human world.
- *Disaster preparedness*. On the practical side prediction is pivotal for coping with a variety of disasters, commonly recognized as major threats to the survival and sustainability of our civilization (e. g. [22]; see also materials of G8-UNESCO World Forum on "Education, Innovation and Research: New Partnership for Sustainable Development", http://g8forum.ictp.it). The reliable advance prediction of extreme events can save lives, contribute to social and economic stability, and to improving the governing of modern societies.

Introduction

Predictability vs. Complexity: The Need for Holistic Approach [7,12,13,15,17,27,32]

Natural science had for many centuries regarded the Universe as a completely predictable machine. As Pierre Simon de Laplace wrote in 1776, "… if we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of the same universe at a succeeding moment." However, at the turn of the 20th century (1905) Jules Henry Poincare discovered, that "… this is not always so. It may happen that small differences in the initial conditions will produce very great ones in the final phenomena. Prediction becomes impossible".

This instability to initial conditions is indeed a definitive attribute of complex systems. Nonetheless, through the robust integral description of such systems, it is possible to discover regular behavior patterns that transcend the inherent complexity. For that reason studying complexity requires the holistic approach that proceeds from the whole to details, as opposed to the reductionism approach that proceeds from details to the whole. It is in principle not possible "to understand a complex system by breaking it apart" [13].

Among the regular behavior patterns of complex systems are "premonitory" ones that emerge more frequently as an extreme event approaches. These premonitory patterns make complex systems predictable. The accuracy of predictions, however, is inevitably limited due to the systems' complexity and observational errors.

Premonitory patterns and extreme events are consecutive manifestations of a system's dynamics. These patterns may not trigger extreme events but merely signal the growth of instability, making the system ripe for the emergence of extreme events.

Methodology

The prediction algorithms described here are based on discovering premonitory patterns. The development of the algorithms requires the integration of complementary methods:

- Theoretical and numerical modeling of complex systems; this includes "universal"models considered in statistical physics and non-linear dynamics (e. g. [1,3,5, 8,12,15,20,42]), and system-specific models, if available.
- Exploratory data analysis.
- Statistical analysis of limited samples, which is relevant since the prediction targets are by definition rare.
- Practical expertise, even if it is intuitive.
- Risk analysis and theory of optimal control for optimizing prediction strategy along with disaster preparedness.

Pattern Recognition of Rare Events This methodology provides an efficient framework for integrating diverse information into prediction algorithms [4,11,19]. This methodology has been developed by the artificial intelligence school of I. Gelfand for the study of rare phenomena of a highly complex origin. In terminology of pattern recognition, the "object of recognition" is the time moment t. The problem is to recognize whether it belongs to the period of alarm, i. e. to a time interval Δ preceding an extreme event. An alarm starts when certain combinations of premonitory patterns emerges.

Several features of that methodology are important for predicting extreme events in the absence of a complete closed theory that would unambiguously define a prediction algorithm. First, this kind of pattern recognition relies on simple, robust parameters that overcome the bane of complexity analysis – incomplete knowledge of the system's causal mechanisms and chronic imperfections in the available data. In its efficient robustness, pattern recognition of rare events is akin to exploratory data analysis as developed by J. Tukey [50]. Second, unlike other statistical methods, e. g. regression analysis, that methodology can be used for small samples such as presidential elections or economic recessions. Also, it integrates quantitative and judgmental parameters and thereby more fully captures the full dimensions of the prediction problem than procedures that rely strictly on quantitative variables.

Summing up, the methodology described here can help in prediction when there are (1) many causal variables, (2) qualitative knowledge about which variables are important, and (3) limited amounts of data [2].

Besides societal predictions, pattern recognition of rare events has been successfully applied in seismology and earthquake prediction (e. g. [11,19,20,44,46]), geological prospecting (e. g. [45]) and in many other fields. Review can be found in [21,47]. Tutorial materials are available at the web site of the Abdus Salam International Centre for Theoretical Physics (http://cdsagenda5.ictp.it/full_display.php?da=a06219).

Validation of Prediction Algorithms The algorithms include many adjustable elements, from selecting the data and defining the prediction targets, to specifying numerical parameters involved. In lieu of theory that would unambiguously determine these elements they have to be developed retrospectively, by "predicting" past extreme events. The application of the methodology to known events creates the danger of self-deceptive data-fitting: As J. von Neumann put it "*with four exponents I can fit an elephant*". The proper validation of the prediction algorithms requires three consecutive tests.

- *Sensitivity analysis*: testing whether predictions are sensitive to variations of adjustable elements.
- *Out of sample analysis*: application of an algorithm to past data that has not been used in the algorithm's development. The test is considered successful if algorithm retains its accuracy.
- *Predicting future events* – the only decisive test of a prediction algorithm (see for example Sect. "Elections" below).

A highly efficient tool for such tests is the error Diagram, showing major characteristics of prediction accuracy [33, 34,35,36,37,38,39]. Its example is given in Fig. 10. Exhaustive sets of these tests are described in [10,11,24,52].

Common Elements of Data Analyzes

The methodology discussed above was used for predicting various kinds of extreme events, as illustrated in the next four Sections. Naturally, from case to case this methodology was used in different ways, according to specifics of phenomena considered. However in all cases data analysis has essential common elements described below.

Sequence of analysis comprises four stages: (i) Defining prediction targets. (ii) Choosing the data (time series), where premonitory patterns will be looked for and summing up a priori constrains on these patterns. (iii) Formulating hypothetical definition of these patterns and developing prediction algorithm; determining the error diagram. (iv) Validating and optimizing that algorithm.

Preliminary transformation of raw data. In predicting recessions (Sect. "US Economic Recessions"), fast acceleration of unemployment (Sect. "Unemployment") and crime surges (Sect. "Homicide Surges") raw data were time series of relevant monthly indicators, hypothetically containing premonitory patterns. Let $f(m)$ be such an indicator, with integer m showing time in months. Premonitory behavior of some indicators is better captured by their linear trends.

Let $W^f(l/q, p)$ be the local linear least-squares regression of a function $f(m)$ within the sliding time window (q, p):

$$W^f(l/q, p) = K^f(q, p)l + B^f(q, p), \quad q \le l \le p\,, \quad (1)$$

where integers l, q, and p stand for time in months.

Premonitory behavior of most indicators was captured by the following two functions:

- The trend of $f(m)$ in the s months long window, $(m - s, m)$. For brevity we denote

$$K^f(m/s) = K^f(m - s, m) \quad (2)$$

- The deviation of $f(m)$ from extrapolation of its long-term regression (i. e. regression on a long time window $(q, m - 1)$:

$$R^f(m/q) = f(m) - W^f(m/q, m - 1)\,. \quad (3)$$

Both functions can be used for prediction since their values do not depend on the information about the future (after the month m) which would be anathema in prediction.

Discretization. The prediction algorithms use one or several premonitory patterns. Each pattern is defined at the lowest – binary – level of resolution, as 0 or 1, distinguishing only the presence of absence of a pattern at each moment of time. Then the objects of recognition are described by binary vectors of the same length. This ensures the robustness of the prediction algorithms.

Simple algorithm called Hamming distance is used for classification of binary vectors in applications considered here, [14,20,28]. Each vector is either premonitory or not. Analyzing the samples of vectors of each class ("the learning material"), the algorithm determines a reference binary vector ("kernel") with components typical for the premonitory vector. Let D be the Hamming distance of a vector from the kernel (the number of non-coinciding binary components). The given vector is recognized as premonitory class, if D is below a certain threshold D^*. This criterion takes advantage of the clustering of precursors in time.

Summing up, these elements of the pattern recognition approach are common for its numerous applications, their diversity notwithstanding. Experience in the specific applications is described in Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges". The conceptual summary of the accumulated experience is given in the final Sect. "Summary: Findings and Emerging Possibilities".

Elections

This Section describes algorithms for predicting the outcome of the US Presidential and mid-term Senatorial elections [28,29,30,31]. Elections' time is set by the law as follows.

- National elections are held every even-numbered year, on the first Tuesday after the first Monday in November (i. e., between November 2 and November 8, inclusively).
- Presidential elections are held once every 4 years, i. e. on every other election day. People in each of the 50 states and District of Columbia are voting separately for "electors" pledged to one or another of the Presidential candidates. These electors make up the "Electoral College" which directly elects the President. Since 1860, when the present two-party system was basically established, the Electoral College reversed the decision of the popular vote only three times, in 1888, 1912, and 2000. Algorithmic prediction of such reversals is not developed so far.
- A third of Senators are elected for a 6-year term every election day; "mid-term" elections held in the middle of a Presidential term are considered here.

Methodology

The prediction target is an electoral defeat of an "incumbent" party, i. e. the party holding the contested seat. Accordingly, the prediction problem is formulated as whether the incumbent party will retain this seat or lose it to the challenging party (*and not whether Republican or Democrat will win*). As is shown below, that formulation is crucial for predicting the outcomes of elections considered.

Data. The pre-election situation is described by robust common sense parameters defined at the lowest (binary) level of resolution, as the *yes* or *no* answers to the questionnaires given below (Tables 1, 2). The questions are formulated in such a way that the answer *no* favors the victory of the challenging party. According to the Hamming distance analysis (Sect. "Common Elements of Data Analyzes") the victory of the challenging party is predicted when the number of answers *no* exceeds a threshold D^*.

Mid-term Senatorial Elections

The prediction algorithm was developed by a retrospective analysis of the data on three elections, 1974, 1978, and 1982. The questionnaire is shown in Table 1. Victory of the challenger is predicted if the number of answers *no* is 5 or more [28,29,30].

The meaning of these questions may be broader than their literal interpretation. For example, financial contributions (key 5 in Table 2) not only provide the resources required for an effective campaign, but may also constitute a poll in which the preferences are weighed by the money attached.

Predicting future elections. This algorithm (without any changes from year to year and from state to state) was applied in advance to the five subsequent elections, 1986–2002. Predictions are shown in Fig. 2. Altogether, 150 seats were put up for election. For each seat a separate prediction was made, 128 predictions were correct, and 22 – wrong.

Statistical significance of this score is 99.9%. In other words the probability to get such a score by chance is below 0.1% [28,29,30]. For some elections these predictions might be considered as trivial, since they coincide with prevailing expectation of experts. Such elections are identified by *Congressional Review*. Eliminating them from the score still results in 99% significance.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 1
Questionnaire for mid-term Senatorial Elections [28]

1.	(Incumbency): The incumbent -party candidate is the sitting senator.
2.	(Stature): The incumbent -party candidate is a major national figure.
3.	(Contest): There was no serious contest for the incumbent -party nomination.
4.	(Party mandate): The incumbent party won the seat with 60% or more of the vote in the previous election.
5.	(Support): The incumbent -party candidate outspends the challenger by 10% or more.
6.	(Obscurity): The challenging -party candidate is not a major national figure or a past or present governor or member of Congress.
7.	(Opposition): The incumbent party is not the party of the President.
8.	(Contest): There is no serious contest for the challenging -party nomination (the nominee gains a majority of the votes cast in the first primary and beats the second-place finisher at least two to one).

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 2
Questionnaire for Presidential elections [29,30]

KEY 1	(Party Mandate): After the midterm elections, the incumbent party holds more seats in the US House of Representatives than it did after the previous midterm elections.
KEY 2	(Contest): There is no serious contest for the incumbent -party nomination.
KEY 3	(Incumbency): The incumbent -party candidate is the sitting president.
KEY 4	(Third party): There is significant third-party or independent campaign.
KEY 5	(Short-term economy): The economy is not in recession during the election campaign.
KEY 6	(Long-term economy): Real per -capita economic growth during the term equals or exceeds mean growth during the previous two terms.
KEY 7	(Policy change): The incumbent administration effects major changes in national policy.
KEY 8	(Social unrest): There is no sustained social unrest during the term.
KEY 9	(Scandal): The incumbent administration is unattained by a major scandal.
KEY 10	(Foreign/military failure): The incumbent administration suffers no major failure in foreign or military affairs.
KEY 11	(Foreign/military success): The incumbent administration achieves a major success in foreign or military affairs.
KEY 12	(Incumbent charisma): The incumbent -party candidate is charismatic or a national hero.
KEY 13	(Challenger charisma): The challenging -party candidate is not charismatic or a national hero.

Presidential Elections

The prediction algorithm was developed by a retrospective analysis of the data on the past 31 elections, 1860–1980; that covers the period between victories of A. Lincoln and R. Reagan inclusively. The questionnaire is shown in Table 2. Victory for the challenger is predicted if the number of answers *no* is 6 or more [29,30].

Predicting of future elections. This algorithm (without any changes from year to year state) was applied in advance to the six subsequent elections, 1984–2004. Predictions are shown in Fig. 3. All of them happened to be correct. In 2000 the decision of popular majority was reversed by the Electoral College; such reversals are not targeted by this algorithm [29,30].

Understanding Elections

Collective behavior. The finding that aggregate-level parameters can reliably anticipate the outcome of both presidential and senatorial elections points to an electoral behavior highly integrated not only for the nation as a whole but also within the diverse American states.

- A presidential election is determined by the collective, integrated estimation of performance of incumbent administration during the previous four years.
- In case of senatorial elections the electorate has more diffused expectations of performance but puts more importance on political experience and status than in the case of presidential elections. Senate incumbents, unlike presidential ones, do not suffer from a bad economy or benefit from a good one. (This suggests that rather than punishing the party holding a Senate seat for hard times, the voters may instead regard the incumbent party as a safe port in a storm).

Similarity. For each election year in all states the outcomes of elections follow the same pattern that transcends the diversities of the situations in each of the individual elections.

The same pattern of the choice of the US President prevails since 1860, i. e. since election of A Lincoln, despite all the overwhelming changes in the electorate, the economy, the social order and the technology of politics during these 130 years. (For example, the electorate of 1860 did not include the groups, which constitute 3/4 of present electorate, such as women, African Americans, or most of the citizens of the Latin American, South European, Eastern European, and Jewish descent [30].

An alternative (and more traditional) concept of American elections focuses on the division of voters into interest and attitudinal groups. By this concept the goal of the contestants is to attract the maximum number of voting blocks with minimal antagonism from other blocks. Electoral choice depends strongly on the factors irrelevant to the essence of the electoral dilemma (e. g. on the campaign tactics). The drawbacks of this concept are discussed in [18,30]. In sum, the work on presidential and senatorial elections described above suggests the following new ways of understanding American politics and perhaps the politics of other societies as well.

1. Fundamental shifts in the composition of the electorate, the technology of campaigning, the prevailing economic and social conditions, and the key issues of campaigns do not necessarily change the pragmatic basis on which voters choose their leaders.
2. It is governing not campaigning that counts in the outcomes of presidential elections.
3. Different factors may decide the outcome of executive as compared to legislative elections.
4. Conventional campaigning will not improve the prospects for candidates faced with an unfavorable combination of fundamental historical factors. Disadvantaged candidates have an incentive to adopt innovative campaigns that break the pattern of conventional politics.
5. All candidates would benefit from using campaigns to build a foundation for governing in the future.

US Economic Recessions

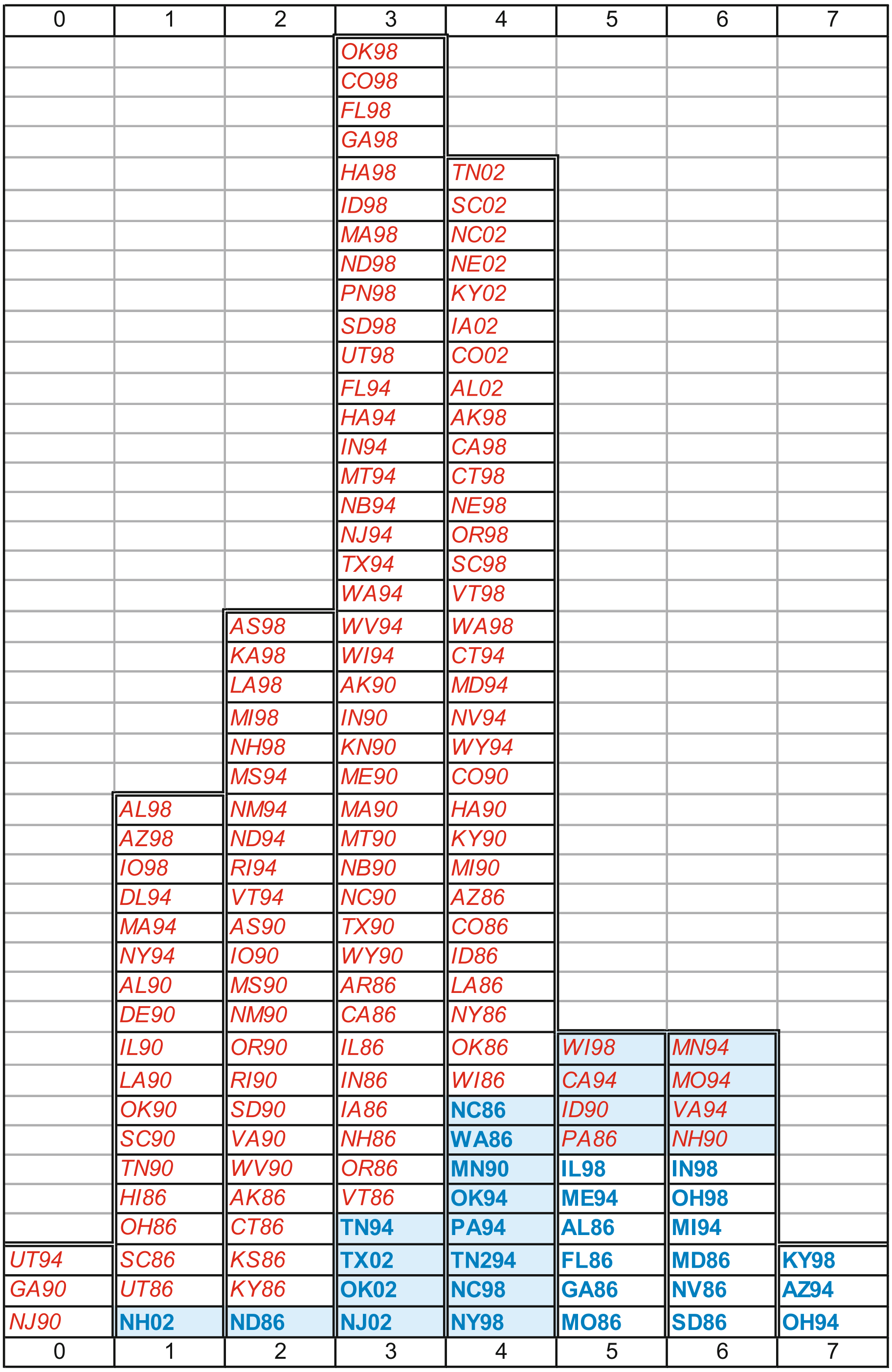
US National Bureau of Economic Research (NBER) has identified the seven recessions that occurred in the US since 1960 (Table 3). The starting points of a recession and of the recovery from it follow the months marked by a peak and a trough of economic activity, respectively.

A peak indicates the last month before a recession, and a trough – the last month of a recession.

Prediction targets considered are the first month after the peak and after the trough ("the turns to the worst and to the best", respectively). The start of the first recession, in 1960, is not among the targets, since the data do not cover a sufficient period of time preceding the recession.

The data used for prediction comprise the following six monthly leading economic indicators obtained from the CITIBASE data base, Jan. 1960–June 2000 (abbreviations are the same, as in [49]).

G10FF = FYGT10 – FEDFUN Difference between the annual interest rate on 10 year US Treasury bonds, and federal fund annual interest rate.

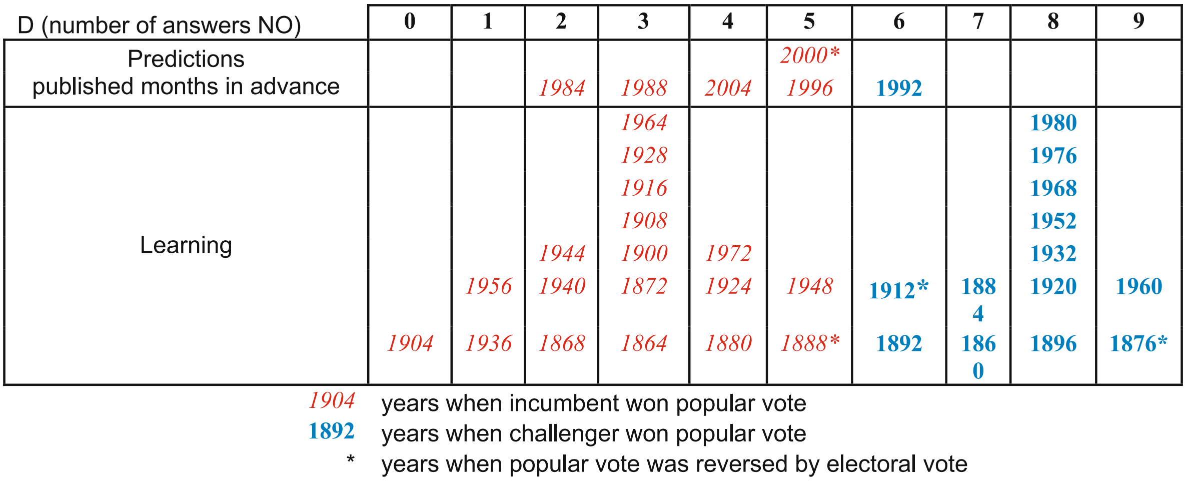


0	1	2	3	4	5	6	7
			OK98				
			CO98				
			FL98				
			GA98				
			HA98	*TN02*			
			ID98	*SC02*			
			MA98	*NC02*			
			ND98	*NE02*			
			PN98	*KY02*			
			SD98	*IA02*			
			UT98	*CO02*			
			FL94	*AL02*			
			HA94	*AK98*			
			IN94	*CA98*			
			MT94	*CT98*			
			NB94	*NE98*			
			NJ94	*OR98*			
			TX94	*SC98*			
			WA94	*VT98*			
		AS98	*WV94*	*WA98*			
		KA98	*WI94*	*CT94*			
		LA98	*AK90*	*MD94*			
		MI98	*IN90*	*NV94*			
		NH98	*KN90*	*WY94*			
		MS94	*ME90*	*CO90*			
	AL98	*NM94*	*MA90*	*HA90*			
	AZ98	*ND94*	*MT90*	*KY90*			
	IO98	*RI94*	*NB90*	*MI90*			
	DL94	*VT94*	*NC90*	*AZ86*			
	MA94	*AS90*	*TX90*	*CO86*			
	NY94	*IO90*	*WY90*	*ID86*			
	AL90	*MS90*	*AR86*	*LA86*			
	DE90	*NM90*	*CA86*	*NY86*			
	IL90	*OR90*	*IL86*	*OK86*	*WI98*	*MN94*	
	LA90	*RI90*	*IN86*	*WI86*	*CA94*	*MO94*	
	OK90	*SD90*	*IA86*	**NC86**	*ID90*	*VA94*	
	SC90	*VA90*	*NH86*	**WA86**	*PA86*	*NH90*	
	TN90	*WV90*	*OR86*	**MN90**	**IL98**	**IN98**	
	HI86	*AK86*	*VT86*	**OK94**	**ME94**	**OH98**	
	OH86	*CT86*	**TN94**	**PA94**	**AL86**	**MI94**	
UT94	*SC86*	*KS86*	**TX02**	**TN294**	**FL86**	**MD86**	**KY98**
GA90	*UT86*	*KY86*	**OK02**	**NC98**	**GA86**	**NV86**	**AZ94**
NJ90	**NH02**	**ND86**	**NJ02**	**NY98**	**MO86**	**SD86**	**OH94**
0	1	2	3	4	5	6	7

OK98 – incumbent won, **KY98** – challenger won, errors are highlighted.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 2

Made-in-advance predictions of the mid-term senatorial elections (1986–2002). Each election is represented by the two-letter state abbreviation with the election year shown by two last digits. Each column shows elections with certain number *D* of answers "*no*" to the questionnaire given in Table 1 (such answers are favorable to challenging party). Value of *D*, indicated at the top, is the Hamming distance from the kernel



D (number of answers NO)	0	1	2	3	4	5	6	7	8	9
Predictions published months in advance			1984	1988	2004	2000* 1996	1992			
Learning	1904	1956 1936	1944 1940 1868	1964 1928 1916 1908 1900 1872 1864	1972 1924 1880	1948 1888*	1912* 1892	188 4 186 0	1980 1976 1968 1952 1932 1920 1896	1960 1876*

1904 years when incumbent won popular vote
1892 years when challenger won popular vote
* years when popular vote was reversed by electoral vote

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 3
Division of presidential elections (1860–2004) by the number *D* of answers "*no*" to the questionnaire given in Table 2 (such answers are favorable to challenging party). *D* is the Hamming distance from the kernel

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 3
American Economic Recessions since 1960

#	Peaks	Troughs
1	1960:04	1961:02
2	1969:12	1970:11
3	1973:11	1975:03
4	1980:01	1980:07
5	1981:07	1982:11
6	1990:07	1991:03
7	2001:03	2001:11

IP Industrial Production, total: index of real (constant dollars, dimensionless) output in the entire economy. This represents mainly the manufacturing industry, because of the difficulties in measuring the quantity of the output in services (such as travel agents, banking, etc.).

LHELL Index of "help wanted" advertising. This is put together by a private publishing company that measures the amount of job advertising (column-inches) in a number of major newspapers.

LUINC Average weekly number of people claiming unemployment insurance.

INVMTQ Total inventories in manufacturing and trade, in real dollars. Includes intermediate inventories (for example held by manufacturers, ready to be sent to retailers) and final goods inventories (goods on the shelves in stores).

FYGM3 Interest rate on 90 day US treasury bills at an annual rate (in percent).

These indicators were already known [48,49], as those that correlate with a recession's approach.

Prediction of a Recession Start

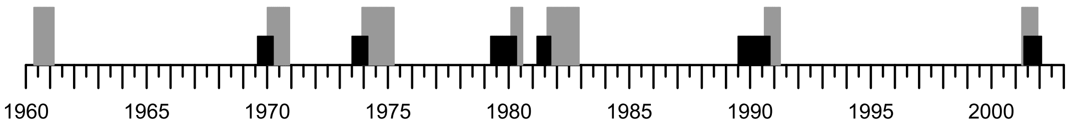
Single indicators exhibit the following premonitory patterns:

G10FF: small value
IP and INVMTQ: small deviation from the long-term trend R^f (3)
FYGM3: large deviation from the long-term trend R^f
LHELL: small trend K^f (2)
LUINC: large trend K^f

The prediction algorithm triggers an alarm after a month when most of the patterns emerge simultaneously. It lasts Δ months and can be extended by the same rule, if premonitory patterns keep emerging. Formal quantitative definition of the algorithm can be found in [23] along with its validation by sensitivity and out-of-sample analyzes.

Alarms and recessions are juxtaposed in Fig. 4. We see that five recessions occurring between 1961 and 2000 were predicted by an alarm. The sixth recession started in April 2001, one month before the corresponding alarm. (Recession of 1960 was not considered for prediction, since data analyzed start just before it.)

Only the first six recessions listed in Table 1 were considered in the developing of the algorithm [23]. Duration



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 4
Alarms (*black bars*) and recessions (*gray bars*)

of each alarm was between 1 and 14 months. Total duration of all alarms was 38 months, or 13.6% of the time interval considered. There were no false alarms. No alarms were yielded so far by subsequent prediction in advance and no recession was identified during that time.

Prediction of a Recession End

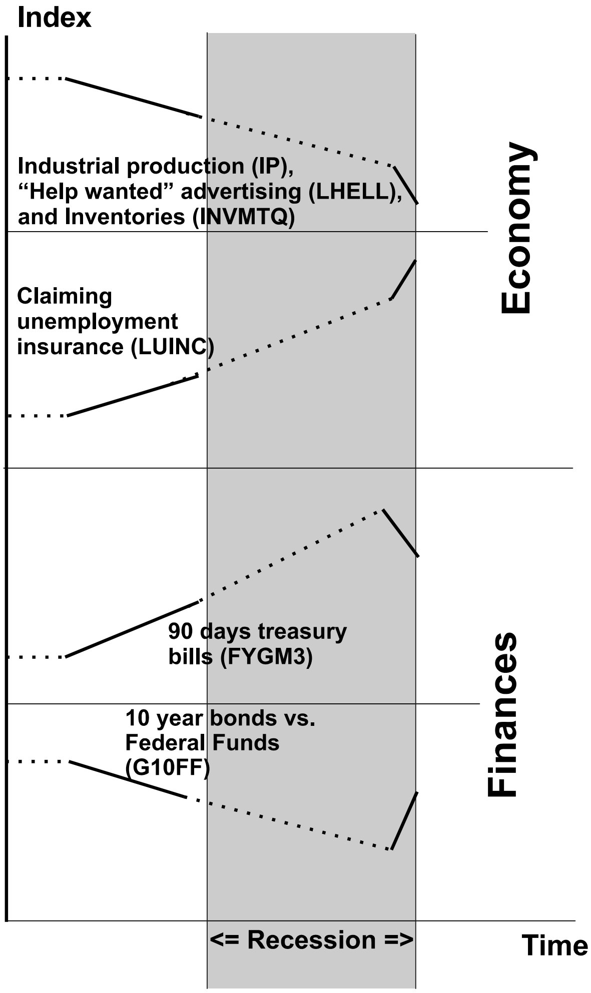
Prediction targets are the starting points of recovery from recessions; these points are indicated in the last column of Table 3.

The data comprise the same six indicators that indicate the approach of a recession (see Subsect. "Prediction of a Recession Start"); they are analyzed only within the recessions' periods.

Data analysis shows intriguing regularity illustrated in Fig. 5:

- Financial indicators change in opposite directions before the recession and before the recovery.
- Economic indicators change in the same direction before the recession and the recovery; but the change is stronger before the recovery, i. e., the economic situation worsens.

Prediction algorithm is formulated in the same terms as in the previous case but an alarm is triggered after *three* consecutive months when most of the patterns emerge simultaneously. The alarms predict when the recovery will start. Alarms and prediction targets are juxtaposed in Fig. 6. Duration of a single alarm is one to five months. Total duration of alarms is 16 months, which is 22% of time covered by all recessions. There are neither false alarms nor failures to predict.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 5
Premonitory changes of indicators before the start of a recession and before its end. See explanations in the text

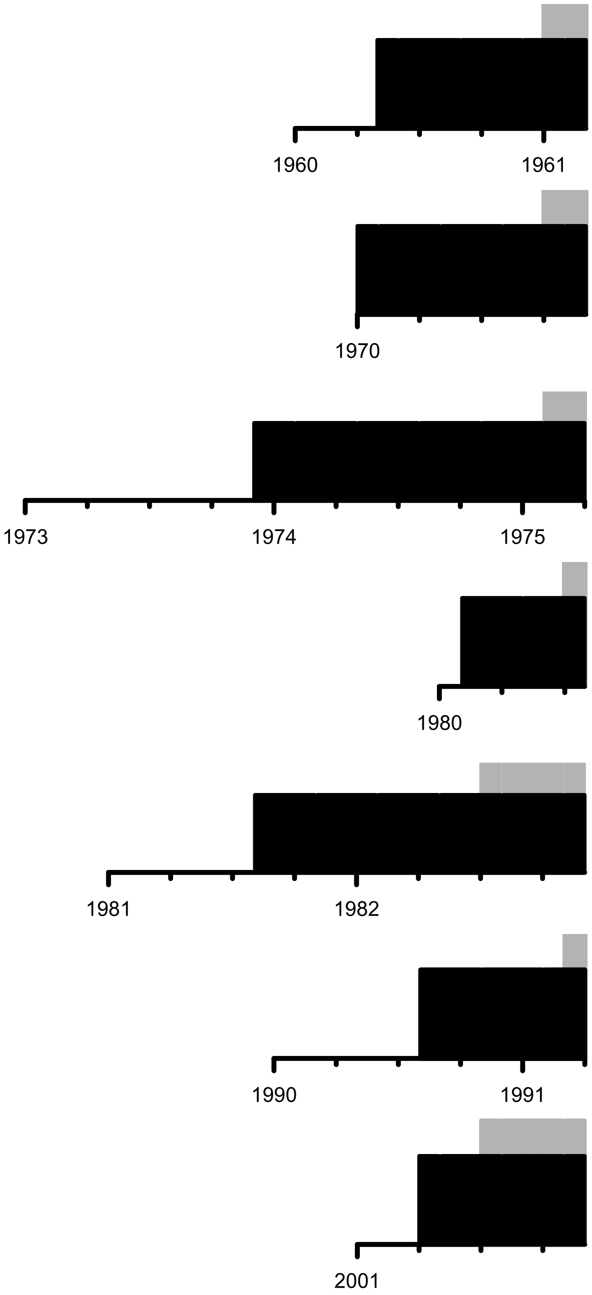
Unemployment

Here we describe uniform prediction of the sharp and lasting unemployment surge in France, Germany, Italy, and the USA [25].

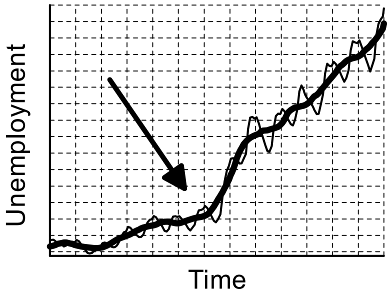
Prediction Target

A prediction target is schematically illustrated in Fig. 7. Thin curve shows monthly unemployment with seasonal variations. On the thick curve seasonal variations are smoothed away. The arrow indicates a sharp upward bend of the smoothed curve. The moment of that bend is the prediction target. It is called by the acronym *FAU*, for "Fast Acceleration of Unemployment".

Smoothing was done as follows: Let $u(m)$ be number of unemployed in a month $m = 1, 2, \dots$. After smooth-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 6
Prediction of recovery from a recession. *Black bars* – periods of recessions. *Gray bars* – alarms preceding the end of a recession



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 7
Fast acceleration of unemployment (*FAU*): schematic definition. *Thin line* – monthly unemployment; with seasonal variations. *Thick line* – monthly unemployment, with seasonal variations smoothed away. The *arrow* indicates a *FAU* – the sharp bend of the smoothed curve. The moment of a *FAU* is the target of prediction

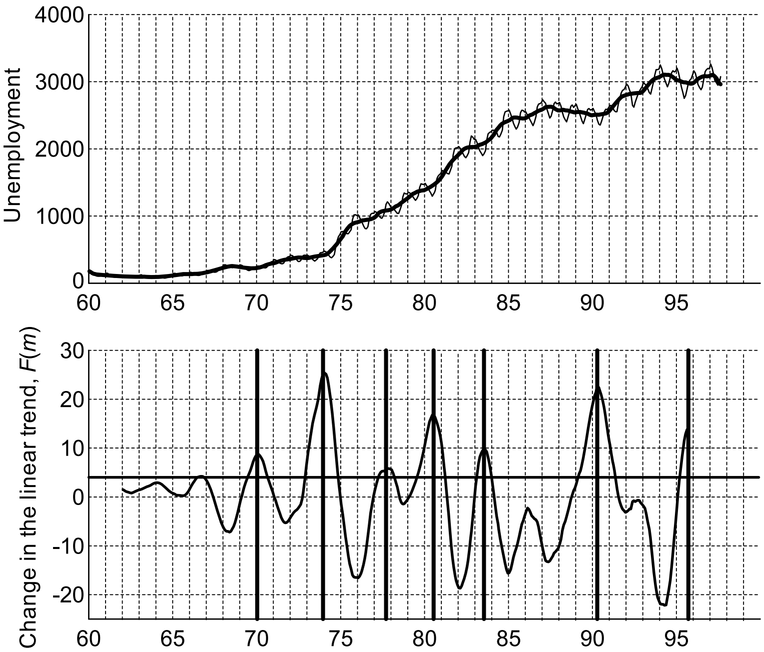
ing out the seasol variation we obtain time series $U(m) = W^u(m/m-6, m+6)$; this is the linear regression over the year-long time interval $(m-6, m+6)$. A natural robust measure of unemployment acceleration at the time m is the bend of the linear trend of U; in notations used in (1) this is the function $F(m/s) = K^U(m+s, m) - K^U(m, m-s)$. The FAUs are identified by the local maxima of $F(m)$ exceeding a certain threshold F. The time m^* and the height F^* of such a maximum are, respectively, the time and the magnitude of a FAU. Subsequent local minimum of $F(m)$ identifies the month m_e when acceleration ends. Figure 8 shows thus defined FAUs for France.

The Data

The analysis has been initially made for France and three groups of data have been analyzed.

- *Composite macroeconomic indicators of national economy*
  1. **IP**: Industrial production indicator, composed of weighted production levels in numerous sectors of the economy, in % relative to the index for 1990.
  2. **L**: Long-term interest rate on 10-year government bonds, in %.
  3. **S**: Short-term interest rate on 3-month bills, in %.
- *Characteristics of more narrow areas of French economy*
  4. **NC**: The number of new passenger car registrations, in thousands of units.
  5. **EI**: Expected prospects for the national industrial sector.
  6. **EP**: Expected prospects for manufacturers.
  7. **EO**: Estimated volume of current orders.

  Indicators 5–7 distinguish only "good" and "bad" expectations determined polling 2,500 manufacturers, whose answers are by the size of their businesses.
- *Indicators related to US economy.*



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 8
Unemployment in France. *Top:* Monthly unemployment, thousands of people. *Thin line*: $u(m)$, data from the OECD database; note the seasonal variations. *Thick line*: $U(m)$, data smoothed over one year. *Bottom:* Determination of *FAU*s. $F(m)$ shows the change in the linear trend of unemployment $U(m)$. *FAU*s are attributed to the local maxima of $F(m)$ exceeding threshold $F = 4.0$ shown by *horizontal line*. The *thick vertical lines* show moments of the *FAU*s

8. **FF/$**: Value of US dollar in French francs.
9. **AR**: The state of the American economy: is it close to a recession or not? This indicator shows the presence or absence of a current pre-recession alarm (see Subsect. "Prediction of a Recession Start").

The data bases with above indicators for Europe are issued by the Organization for Economic Cooperation and Development [43] and the International Monetary Fund [16].

American analogues of indicators **IP**, **L**, and **S** are provided by CITIBASE; they are described in Sect. "US Economic Recessions>" under abbreviations **IP**, **FYGM3** and **FIGT10** respectively.

Prediction

Single indicators exhibit the following premonitory behavior.

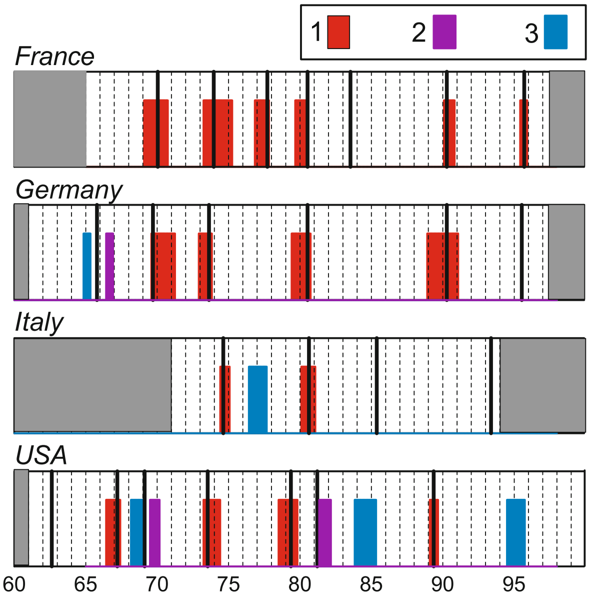
- Steep upward trends of composite indicators (#1–#3). This behavior reflects "overheating" of the economy and may sound counterintuitive for industrial production (#1), since the rise of production is supposed to create more jobs. However, a particularly steep rise may create oversupply.
- Steep downward trends of economic expectations by general public (#4) and business community (#5–#8).
- Proximity of an American recession (#9). Before analysis was made such and opposite precursors might be expected for equally plausible reasons, so that this finding, if further confirmed, does provide a constraint on understanding unemployment's dynamics.

Among different combinations of indicators the macroeconomic ones (#1–#3) jointly give relatively better predictions, with smallest rates of errors and highest stability in sensitivity tests.

Retrospective prediction. Macroeconomic indicators were used jointly in the Hamming distance prediction algorithm (Sect. "Common Elements of Data Analyzes"). Being robust and self-adjusting to regional conditions, this algorithm was applied without any changes to the four countries considered here.

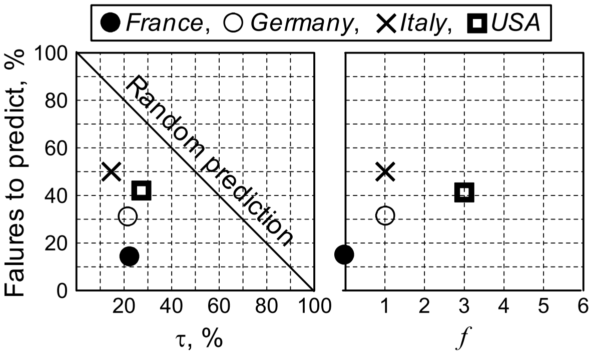
Alarms and *FAU*s are juxtaposed in Fig. 9. Error diagram in Fig. 10 shows quality of prediction for different countries. For US the quality is lower than for European countries, though still higher than in random predictions.

Prediction of the future FAUs was launched for USA. The results are shown in Fig. 11. It shows that by Jan-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 9

Retrospective predictions for four countries: *FAUs* and alarms obtained by the prediction algorithm. The *thick vertical lines* show the moments of *FAUs* in a country. *Bars* – the alarms with different outcome: 1 – alarms that predict *FAUs*, 2 – alarms starting shortly after *FAUs* within the periods of unemployment surge, 3 – false alarms. *Shaded areas* on both sides indicate the times, for which data on economic indicators were unavailable



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 10

Error diagram for prediction of *FAUs* in different countries; τ is total duration of alarms in % to the time interval considered, *f* – total number of false alarms

uary 2008 two correct predictions have been made, without ether false alarms or failures to predict. In November 2006 the second prediction was filed on the web site of the Anderson School of Management, University of California, Los Angeles (http://www.uclaforecast.com/). This started the documented experiment in testing the algorithm by predicting future *FAU*s on that website.

Homicide Surges

This section analyzes the prediction of homicide rates in an American megacity – Los Angeles, CA [24].

Prediction Target

A prediction target is the start of a sharp and lasting acceleration of the homicide rate; it is called by the acronym SHS, for "Start of the Homicide Surge." It is formally determined by the analysis of monthly homicides rates, with seasonal variations smoothed out, as described in Subsect. "Prediction Target". Prediction targets thus identified are shown by vertical lines in Figs. 12 and 14 below.

The Data

The analyzed data include monthly rates of the homicides and 11 types of lesser crimes, listed in Table 2. Definitions of these crimes are given in [6].

The data are taken from two sources:

- The National Archive of Criminal Justice Data, placed on the web site (NACJD), 1975–1993.
- Data bank of the Los Angeles Police Department(LAPD) Information Technology Division), 1990–2003.

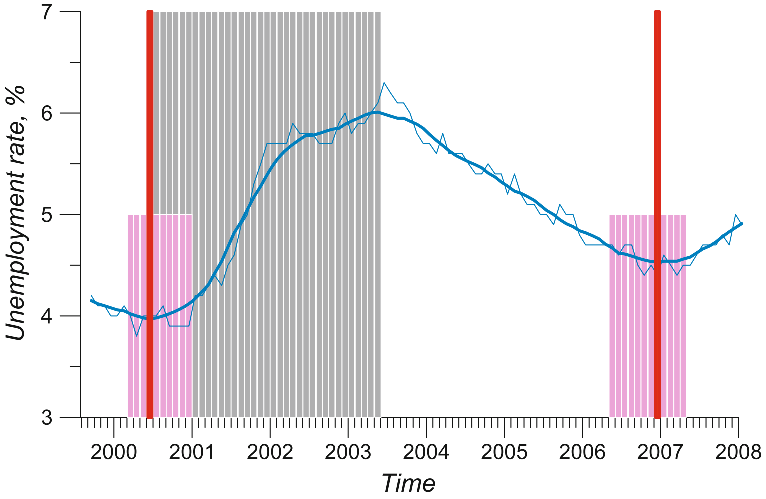
The algorithm does not use socio-economic determinants of crime, or other data that might be also useful. The objective was to develop a simple, efficient prediction model; development of comprehensive causal model would be a complementary objective.

Prediction

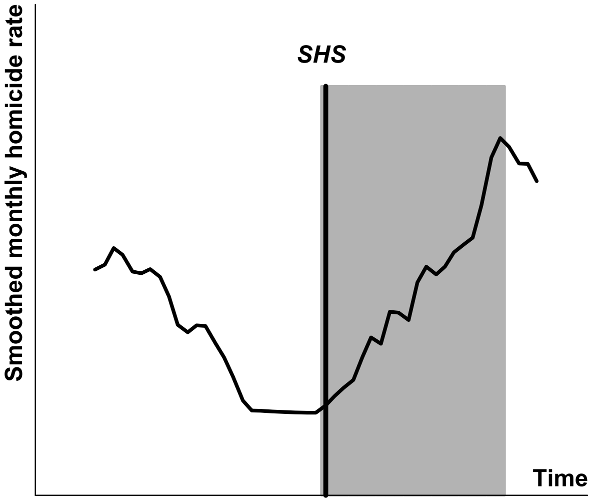
Premonitory behavior of indicators is illustrated in Fig. 13. The first phase is characterized by an escalation of burglaries and assaults, but not of robberies. Later on, closer to a homicide surge, robberies also increase.

The Prediction algorithm based on Hamming distance (see Sect. "Common Elements of Data Analyzes") uses seven indicators listed in Table 4. Other five indicators marked by * are used in sensitivity tests; and the homicide rate is used for identification of targets *SHS*.

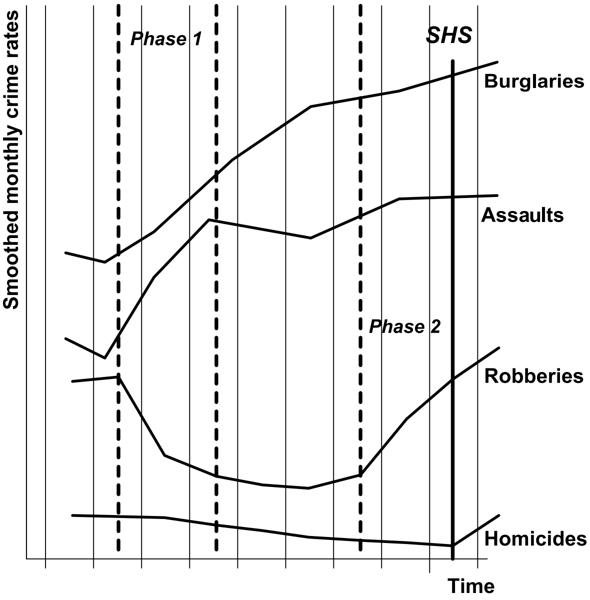
Alarms and homicide surges are juxtaposed in Fig. 14. The *SHS* episode in November 1994 has occurred simultaneously with the corresponding alarm. It is captured by an alarm, which starts in the month of *SHS* without a lead time. Prediction missed the October 1999 episode: it occurred two months *before* the start of the corresponding



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 11
Experiment in predicting future FAUs, September (1999)–January (2008). ***Thin blue curve*** **shows monthly unemployment rate in USA, according to data of Bureau of Labor Statistics, US Department of Labor (http://www.data.bls.gov).** ***Thick curve*** **shows this rate with seasonal variation smoothed away.** ***Vertical red lines*** **show prediction targets – the moments of** ***FAU***, ***gray bar*** **– the period of unemployment's growth;** ***pink bars*** **– periods of alarms**



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 12
Target of prediction – the Start of the Homicide Surge ("*SHS*"); schematic definition. ***Gray bar*** **marks the period of homicide surge**



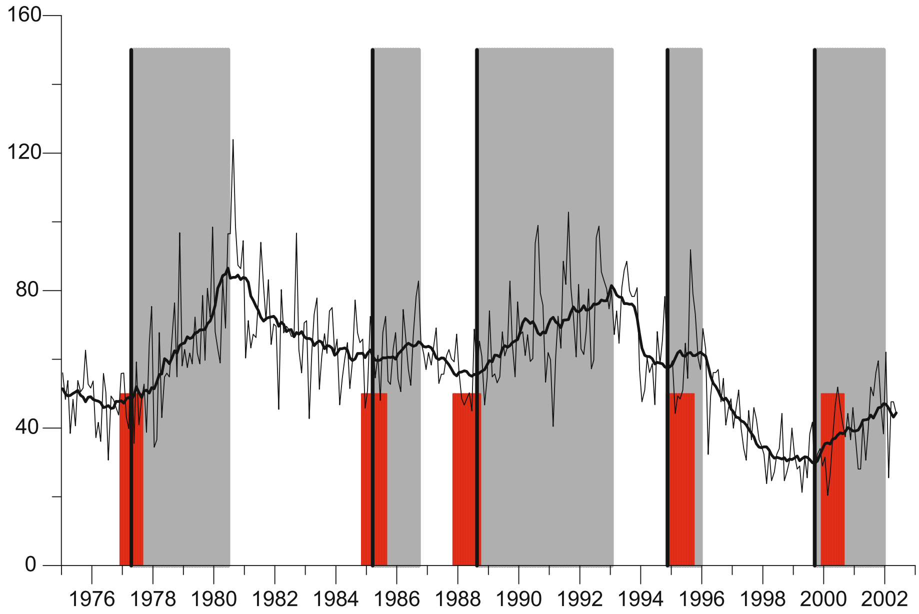
Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 13
Scheme of premonitory changes in crime statistics

alarm. Such delays should be taken into account for validating the algorithm. Note, however, that the last prediction did remain informative.

Altogether alarms occupy 15% of the time considered. During phase 2 (as defined in Fig. 13) this rate might be reduced [24].

Summary: Findings and Emerging Possibilities

The findings described above enhance predictive understanding of extreme events and indicate yet untapped possibilities for further R&D in that field.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 14
Performance of prediction algorithm through 1975–2002. *Thin curve* – original time series, total monthly number of homicides in Los Angeles city, per 3,000,000 inhabitants. Data from NACJD [6] have been used for 1975–1993 and from the Data Bank of the Los Angeles Police Department (LAPD Information Technology Division) for subsequent 9 years. *Thick curve* – smoothed series, with seasonal variations eliminated. *Vertical lines* show the targets of prediction – episodes of *SHS* (Subsect. "Prediction Target"). *Gray bars* show the periods of homicide surge. *Red bars* show the alarms declared by the prediction algorithm [24]

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 4
Types of crimes considered (after [6])

Homicide	Robberies	Assaults	Burglaries
● All	● All	● All*	● Unlawful not forcible entry
	● With firearms	● With firearms	● Attempted forcible entry*
	● With knife or cutting instrument	● With knife or cutting instrument	
	● With other dangerous weapon	● With other dangerous weapon*	
	● Strong -arm robberies*	● Aggravated injury assaults*	

*Analyzed in sensitivity tests only

Pattern Recognition Approach

Information extracted from the already available data is indeed increased by this approach. To each problem considered here one may apply the following conclusion of J. Stock, a leading expert in the field: "Prediction/of recessions/requires fitting non-linear, high-dimensional models to a handful of observations generated by a possibly non-stationary economic environment.... The evidence presented here suggests that these simple binary transformations of economic indicators have significant predictive content for recessions. It is striking that these models, in which the information in the data is reduced to binary indicators, have predictive contents comparable to or, in many cases, better than that of more conventional models." Importantly, this is achieved by using not more detailed data and models, but more robust aggregation (Subsect. "Predictability vs. Complexity: The Need for Holistic Approach").

Partial "universality" of premonitory patterns is established by broad research in modeling and data analysis. This includes the common definition of the patterns, their self-adjustment, scaling, and similarity [9,10,20,26,42]; see also references in Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges").

Relation to "cause and effect" analysis (*perpetrators or witnesses?*). Premonitory patterns might be either "perpetrators" contributing to causing extreme events, or the "witnesses" – parallel manifestations of the system's development. The cause that triggered a specific extreme event is usually identified, at least in retrospect. It may be, for example, a certain governmental decision, a change in the international situation, a natural disaster, the depletion of natural resources etc. However an actual extreme event might materialize only if the system is destabilized and "ripe" for it. Patterns of each kind signal such a ripe situation.

What premonitory patters to use for prediction? Existing theories and experience reduce the number of such patterns, but too many of them remain hypothetically promising and have to be chosen by a trial and error procedure. Inevitably a prediction algorithm begins with a limited number of promising patterns. They should be sufficient for prediction, but other patterns may be equally or more useful and should be considered in further development of the algorithm. Most relevant "perpetrators" might not be included in the most useful patterns (e. g. due to their sensitivity to too many factors).

Relation to policy-making: *prediction and disaster preparedness*. Reliable predictions of future extreme events in complex societal systems would allow policy-makers to take remedial action before rather than after the onset of such afflictions as economic disasters, crime surges, etc. As in case of military intelligence predictions would be useful if their accuracy is known, albeit not necessarily high. Analysis of error diagrams allows to regulate the tradeoff between the rates of failures to predict and false alarms according to the needs of a decision-maker.

Relation to governing and campaigning. The findings presented here for the USA elections show that top elected officials would have better chances for reelection, if they focus on effective governing, and not on rhetoric, packaging and image-making. Candidates will benefit themselfes and their parties if they run substantive campaigns that build a foundation for governing during the next term.

Further Possibilities

A wealth of *yet untapped data and models* is readily available for the continuation of the kinds of studies described and analyzed in this article. Following are some immediate possibilities; specific examples can be found in the given references.

- *Continuing experiments in advance prediction*, for which the above findings set up a base (Sect. "Elections"). Successes and errors are equally important [37,38].
- *Incorporating other available data into the analysis* (Sects. "US Economic Recessions","Unemployment")
- *Predicting the same kind of extreme events in different contexts* (Sect. "Unemployment")
- *Predicting the end of a crisis* (Sect. "US Economic Recessions").
- *Multistage prediction with several lead times* (Sect. "Homicide Surges")

  Less imminent, but within reach are:
- *"Universal" scenarios of extreme development and low-parametric definition of an ensemble of premonitory patterns* [9,51,52].
- *Validation of an algorithm and joint optimization of prediction and preparedness strategy* [38].
- *Developing prediction algorithms for other types of extreme events.*

The authors would be glad to provide specific information upon request.

Generalizations

The problems considered here have the following common features:

- *The absence of a closed theory* that would unambiguously determine prediction methodology. This leads to the need for intense intertwining of mathematics, statistical physics and non-linear dynamics, a range of societal sciences, and practical experience (Subsect. "Methodology"). In reality this requires long-term collaboration of respective experts. As can be seen from the references to Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges" previous applications inevitably involved the teams of such experts.
- *Predictions in advance* is the only final validation of the results obtained.
- *The need for holistic analysis* driven to extreme robustness.
- *Considerable, albeit limited, universality* of the premonitory phenomena.

Two classical quotations shed the light on these features:

A. N. Kolmogoroff. "It became clear for me that it is unrealistic to have a hope for the creation of a pure theory [of the turbulent flows of fluids and gases] closed in itself. Due to the absence of such a theory we have to rely upon the hypotheses obtained by processing of the experimental data."

M. Gell-Mann: "… if the parts of a complex system or the various aspects of a complex situation, all defined in advance, are studied carefully by experts on those parts or aspects, and the results of their work are pooled, an adequate description of the whole system or situation does not usually emerge. … The reason, of course, is that these parts or aspects are typically entangled with one another. … We have to supplement the partial studies with a transdisciplinary crude look at the whole."

In the general scheme of things the problem considered belongs to a much wider field – the quest for a universal theory of complex systems extended to predicting extreme events – the Holy Grail of complexity studies. This quest encompasses the natural and human-made complex systems that comprise what some analysts have called "the global village". It requires entirely new applications of modern science, such as algebraic geometry, combinatorics, and thermodynamics. As a means for anticipating, preventing and responding to natural and manmade disasters and for improving the outcomes of economic and political systems, the methods described here may hold one key for the survival and sustainability of our civilization.

Bibliography

Primary Literature

1. Allègre CJ, Le Mouël J-L, Ha Duyen C, Narteau C (1995) Scaling organization of fracture tectonics (SOFT) and earthquake mechanism. Phys Earth Planet Inter 92:215–233
2. Armstrong JS, Cuzan AG (2005) Index methods for forecasting: An application to american presidential elections. Foresight Int J Appl Forecast 3:10–13
3. Blanter EM, Shnirman MG, Le Mouël JL, Allègre CJ (1997) Scaling laws in blocks dynamics and dynamic self-organized criticality. Phys Earth Planet Inter 99:295–307
4. Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning prog in the detection of oil containing layers. Geol Geofiz 6:96–105
5. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seismol Soc Am 57:341–360
6. Carlson SM (1998) Uniform crime reports: Monthly weapon-specific crime and arrest time series 1975–1993 (National, State, 12-City Data), ICPSR 6792 Inter-university Consortium for Political and Social Research. Ann Arbor
7. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. Phys Rev Lett 59:845
8. Gabrielov A, Keilis-Borok V, Zaliapin I, Newman WI (2000) Critical transitions in colliding cascades. Phys Rev E 62:237–249
9. Gabrielov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. arXiv:0708.1542
10. Gabrielov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. Geophys J Int 143(2):427–437
11. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. Phys Earth Planet Inter 11:227–283
12. Gell-Mann M (1994) The quark and the jaguar: Adventures in the simple and the complex. Freeman, New York
13. Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) Chaos Sci Am 255:46–57
14. Gvishiani AD, Kosobokov VG (1981) On found of the pattern recognition results applied to earthquake-prone areas. Izvestiya Acad Sci USSR. Phys Earth 2:21–36
15. Holland JH (1995) Hidden order: How adaptation builds complexity. Addison, Reading
16. IMF (1997) International monetary fund, international financial statistics. CD-ROM
17. Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) Phase transitions and critical phenomena, vol 5a. Academic, London, pp 1–34
18. Keilis-Borok VI, Lichtman AJ (1993) The self-organization of American society in presidential and senatorial elections. In: Kravtsov YA (ed) Limits of predictability. Springer, Berlin, pp 223–237
19. Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allègre CJ (ed) Source mechanism and earthquake prediction applications. Editions du centre national du la recherché scientifique, Paris, pp 51–60
20. Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Berlin
21. Keilis-Borok V, Soloviev A (2007) Pattern recognition methods and algorithms. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-11
22. Keilis-Borok VI, Sorondo MS (2000) (eds) Science for survival and sustainable development. The proceedings of the study-week of the Pontifical Academy of Sciences, 12–16 March 1999. Pontificiae Academiae Scientiarvm Scripta Varia, Vatican City
23. Keilis-Borok V, Stock JH, Soloviev A, Mikhalev P (2000) Pre-recession pattern of six economic indicators in the USA. J Forecast 19:65–80
24. Keilis-Borok VI, Gascon DJ, Soloviev AA, Intriligator MD, Pichardo R, Winberg FE (2003) On predictability of homicide surges in megacities. In: Beer T, Ismail-Zadeh A (eds) Risk science and sustainability. Kluwer, Dordrecht (NATO Sci Ser II Math, Phys Chem 112), pp 91–110
25. Keilis-Borok VI, Soloviev AA, Allègre CB, Sobolevskii AN, Intriligator MD (2005) Patterns of macroeconomic indicators preceding the unemployment rise in Western Europe and the USA. Pattern Recogn 38(3):423–435
26. Keilis-Borok V, Soloviev A, Gabrielov A, Zaliapin I (2007) Change of scaling before extreme events in complex systems. In: Proceedings of the plenary session on "predictability in science: Accuracy and limitations", Pontificiae Academiae Scientiarvm Scripta Varia, Vatican City
27. Kravtsov YA (ed) (1993) Limits of predictability. Springer, Berlin
28. Lichtman AJ, Keilis-Borok VI (1989) Aggregate-level analysis and prediction of midterm senatorial elections in the United States, 1974–1986. Proc Natl Acad Sci USA 86(24):10176–10180
29. Lichtman AJ (1996) The keys to the White House. Madison Books, Lanham
31. Lichtman AJ (2005) The keys to the White House: Forecast for 2008. Foresight Int J Appl Forecast 3:5–9
30. Lichtman AJ (2008) The keys to the White House, 2008 edn. Rowman/Littlefield, Lanham
32. Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) Earthquake prediction: Nine major earthquakes in china. Springer, New York
33. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) Forecast verification. A practitioner's guide in atmospheric science. Wiley, Chichester, pp 37–76

34. Molchan GM (1990) Strategies in strong earthquake prediction. Phys Earth Planet Inter 61:84–98
35. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. Tectonophysics 193:267–276
36. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) Computational seismology and geodynamics, vol 1. Am Geophys Un, Washington, pp 1–10
37. Molchan GM (1997) Earthquake prediction as a decision-making problem. Pure Appl Geophys 149:233–237
38. Molchan GM (2003) Earthquake prediction strategies: A theoretical analysis. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Berlin, pp 209–237
39. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. Geophys J Int 173(3):1012–1017
40. NACJD: http://www.icpsr.umich.edu/NACJD/index.html
41. NBER: http://www.nber.org/cycles/cyclesmain.html
42. Newman W, Gabrielov A, Turcotte DL (eds) (1994) Nonlinear dynamics and predictability of geophysical phenomena. Am Geophys Un, Int Un Geodesy Geophys, Washington
43. OECD (1997) Main economic indicators: Historical statistics 1960–1996. Paris, CD-ROM
44. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. Nature 256:270–273, London
45. Press F, Briggs P (1977) Pattern recognition applied to uranium prospecting. Nature 268:125–127
46. Press F, Allen C (1995) Patterns of seismic release in the southern California region. J Geophys Res 100(B4):6421–6430
47. Soloviev A (2007) Application of the pattern recognition techniques to earthquake-prone areas determination. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-9
48. Stock JH, Watson MW (1989) New indexes of leading and coincident economic indicators. NBER Macroecon Ann 4:351–394
49. Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators. In: Stock JH, Watson MW (eds) Business cycles, indicators, and forecasting (NBER Studies in Business Cycles, vol 28), pp 95–156
50. Tukey JW (1977) Exploratory data analysis. Addison-wesley series in behavioral science: Quantitative methods. Addison, Reading
51. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Geocomplexity and the physics of earthquakes. Am Geophys Un, Washington
52. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades, II: Prediction of critical transitions. J Stat Phys 111(3–4):839–861

Books and Reviews

Bongard MM (1967) The problem of recognition. Nauka, Moscow

Brito DL, Intriligator MD, Worth ER (1998) In: Eliassson G, Green C (eds) Microfoundations of economic growth: A Schumpeterian perspective. University of Michigan Press, Ann Arbor

Bui Trong L (2003) Risk of collective youth violence in french suburbs. A clinical scale of evaluation, an alert system. In: Beer T, Ismail-Zadeh A (eds) Risk science and sustainability. Kluwer, Dordrecht (NATO Sci Ser II Math Phys Chem 112)

Engle RF, McFadden DL (1994) (eds) Handbook of econometrics, vol 4. North-Holland, Amsterdam

Klein PA, Niemira MP (1994) Forecasting financial and economic cycles. Wiley, New York

Messner SF (1983) Regional differences in the economic correlates of the urban homicide rate. Criminology 21:477–488

Mitchell WC (1951) What happens during business cycles: A progress report. NBER, New York

Mitchell WC, Burns AF (1946) Measuring business cycles. NBER, New York

Moore GH (ed) (1961) Business cycle indicators. NBER, New York

Mostaghimi M, Rezayat F (1996) Probability forecast of a downturn in US economy using classical statistical theory. Empir Econ 21:255–279

Watson MW (1994) In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol IV. North-Holland, Amsterdam

Finance and Econometrics, Introduction to

BRUCE MIZRACH
Department of Economics, Rutgers University,
New Jersey, USA

Article Outline

Introduction

Economics and finance have slowly emerged from the Walrasian, representative agent paradigm exemplified by the research agenda in general equilibrium theory. This program may have reached its pinnacle in the 1970s, with a highly abstract treatment of the existence of a market clearing mechanism. The normative foundation of this research was provided by powerful welfare theorems that demonstrated the optimality of the market allocations. Unfortunately, this abstract world had little economics in it. The models rarely provided empirical implications. Lifetime consumption and portfolio allocation plans were formed in infancy, unemployment was Pareto optimal, and the role for government was largely limited to public goods provision.

The demonstration by Benhabib, Brock, Day, Gale, Grandmont, [1,4,8,9] and others, that even simple mathematical models could display highly complex dynamics was the beginning of a new research program in economics. This section on finance and econometrics surveys some of the developments of the last 20 years that were inspired by this research.

Econometrics

Time series econometrics was originally built on the representation theorems for Euclidean spaces. The existence of a Wold decomposition in linear time series led to the widespread use of Box–Jenkins [3] style modeling as an alternative to structural or reduced form models.

A number of stylized facts about the economy emerged that simply could not be explained in this linear world. Rob Engle [2] and Tim Bollerslev [5] showed that volatility was quite persistent, even in markets that appeared to be nearly random walks. In ► GARCH Modeling, Christian Hafner surveys the extensive development in this area.

James Hamilton [10] and Salih Neftci [11] demonstrated that the business cycle was asymmetric and could be well described by a Markov switching model. James Morley ► Macroeconomics, Non-linear Time Series in and Jeremy Piger ► Econometrics: Models of Regime Changes describe the developments in this area. Virtually all the moments, not just the conditional mean, are now thought to be varying over the business cycle. These models help us to understand why recessions are shorter than expansions and why certain variables lead and lag the cycle.

Nearly all the business cycle models involve the use of latent or unobservable state variables. This reflects a reality that policy makers themselves face. We rarely know whether we are in a recession until it is nearly over. These latent variable models are often better described in a Bayesian rather than a classical paradigm. Oleg Korenok ► Bayesian Methods in Non-linear Time Series provides an introduction to the frontier research in this area.

Markets are often drawn towards equilibrium states in the absence of exogenous shocks, and, since the 1940s, this simple idea was reflected in the building of macroeconometric models. In linear models, Engle and Granger [6] formalized this notion in an error correction framework. When the adjustment process is taking place between two variables that are not stationary, we say that they are cointegrated. Escanciano and Escribano extend the error correction framework and cointegration analysis to nonlinear models in ► Econometrics: Non-linear Cointegration.

Because we often know very little about the data generating mechanism for an economy, nonparametric methods have become increasingly popular in the analysis of time series. Cees Diks discusses in ► Nonparametric Tests for Independence methods to analyze both data and the residuals from an econometric model.

Our last two entries look at the data generated by individual consumers and households. Pravan Trivedi ► Microeconometrics surveys the microeconometric literature, and Jeff Wooldridge ► Econometrics: Panel Data Methods examines the tools and techniques useful for analyzing cross-sectional data.

Agent Based Modeling

The neo-classical synthesis in economics was built upon the abstraction of a single optimizing agent. This assumption simplified the model building and allowed for analytical solutions of the standard models. As computational power became cheaper, it became easier to relax these assumptions. Many economists underestimated the complexity of a world in which multiple agents interact in a dynamic setting. Econophysicists, as Bertrand Roehner describes in ▶ Econophysics, Observational, were not surprised. Roehner is just one of scores of physicists who have brought their tools and perspectives to economics.

Agent based modeling has had a large impact on finance. Financial economics had been led by a Chicago influenced school that saw markets as both rational and efficient. Behavioral finance has eroded the view that people always make optimizing decisions even when large sums of money are at stake. The boundedly rational agents in Sebastiano Manzan's ▶ Finance, Agent Based Modeling in are prone to speculative bubbles. Markets crash suddenly in agent based computational models and in large scale experimental stock markets.

Finance

The foundation of financial economics is the theory of optimal consumption and saving. The goal of the empirical literature was to identify a set of risk factors that would explain why certain assets have a higher return than others. Ralitsa Petkova ▶ Financial Economics, The Cross–Section of Stock Returns and the Fama-French Three Factor Model surveys the canonical model of Fama and French [7] and the extensions to this model in the last decade.

With risk averse agents, asset returns are often predictable. Stijn van Nieuwerburgh and Ralph S.J. Koijen ▶ Financial Economics, Return Predictability and Market Efficiency demonstrate the robustness of this result in a structural model and show that the dividend price ratio does predict future stock returns.

Mototsugu Shintani addresses in ▶ Financial Forecasting, Sensitive Dependence the concept of predictability from an information theoretic perspective through the use of Lyapunov exponents. The exponents not only tell us which systems display sensitive dependence on initial conditions ("chaos") but also provide a predictive horizon for data generated by the model. Shintani finds that financial data appear to not be chaotic, even though they display local dependence on initial conditions.

Mark Kamstra and Lisa Kramer's entry on ▶ Financial Economics, Time Variation in the Market Return primarily focus on the equity premium, the substantially higher return in the US and other countries on equities, over default free securities like Treasury bonds. They document its statistical significance and discuss some behavioral explanations. They demonstrate that behavioral moods can influence asset prices.

Terence Mills' ▶ Financial Economics, Non-linear Time Series in surveys the use of nonlinear time series techniques in finance. Gloria Gonzalez-Rivera and Tae-Hwy Lee look at the ability of nonlinear models to forecast in ▶ Financial Forecasting, Non-linear Time Series in. They also cover the methodology for assessing forecast improvement. The best forecast may not be the one that predicts the mean most accurately; it may instead be the one that keeps you from large losses.

Our last two papers in this area focus on volatility. Markus Haas and Christian Pigorsch discuss the ubiquitous phenomenon of fat-tailed distributions in asset markets in ▶ Financial Economics, Fat-Tailed Distributions. They provide evidence on the frequency of extreme events in many different markets, and develop the implications for risk management when the world is not normally distributed. Torben Andersen and Luca Benzoni ▶ Stochastic Volatility introduce the standard volatility model from the continuous time finance literature. They contrast it with the GARCH model discussed earlier and develop econometric methods for estimating volatility from discretely sampled data.

Market Microstructure

Market microstructure examines the institutional mechanisms by which prices adjust to their fundamental values. The literature has grown with the availability of transactions frequency databases. Clara Vega and Christian Miller ▶ Market Microstructure survey the topic largely from a theoretical perspective. Because disparate markets are likely to have different mechanisms and regulators, the literature has evolved by instrument. Carol Osler ▶ Market Microstructure, Foreign Exchange examines the microstructure of the foreign currency market, the largest and most liquid asset market. Bruce Mizrach and Chris Neely ▶ Treasury Market, Microstructure of the U.S. look at the government bond market in the US as it has evolved into an electronic market. Michael Piwowar ▶ Corporate and Municipal Bond Market Microstructure in the U.S. looks at two bond markets with a large number of issues that trade only very infrequently. Both the markets which he examines have become substantially more transparent through recent government initiatives.

Conclusion

This section covers a wide range of material from theoretical time series analysis to descriptive modeling of financial markets. The theme of complexity is a unifying one in the sense that the models are generally nonlinear and can produce a wide range of possible outcomes. There is complexity in the data which now evolves at a millisecond frequency. Readers should find a variety of perspectives and directions for future research in a heterogenous but interconnected range of fields.

Acknowledgments

I would like to thank all of the contributors to this section of the encyclopedia.

Bibliography

1. Benhabib J, Day RH (1982) A characterization of erratic dynamics in the overlapping generations models. J Econ Dyn Control 4:37–55
2. Bollerslev TP (1986) Generalized autoregressive conditional heteroscedasticity. J Econ 31:307–327
3. Box G, Jenkins G (1994) Time Series Analysis Forecasting and Control, 3rd ed. Prentice Hall
4. Brock WA (1986) Distinguishing random and deterministic systems. J Econ Theory 40:168–195
5. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. Econ 50:987–1008
6. Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation, and testing. Econometrica 55:251–276
7. Fama E, French K (1992) The cross-section of expected stock returns. J Finance 47:427–465
8. Gale D (1973) Pure exchange equilibrium of dynamic economic models. J Econ Theory 6:12–36
9. Grandmont JM (1985) On Endogenous Competitive Business Cycles. Econometrica 53:995–1045
10. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
11. Neftçi SH (1984) Are economic time series asymmetric over the business cycle? J Political Econ 92:307–328

Finance, Agent Based Modeling in

SEBASTIANO MANZAN
Department of Economics and Finance, Baruch College
CUNY, New York, USA

Article Outline

Glossary

Rational expectations (RE) An assumption often introduced in economic models. It assumes that agents subjective distribution is equal to the true probability distribution of a random variable. The implication is that expectation errors are purely random.

Bounded rationality The assumption that agents have limited ability to acquire and process information and to solve complex economic problems. These limitations imply that expectations can diverge from RE.

Efficient markets hypothesis (EMH) An application of rational expectations to asset prices. The EMH assumes that asset prices reflect all available information. It implies that asset prices behave like a random walk process and their changes are purely random.

Artificial financial markets A market populated by agents that have bounded rational expectations and learning from available information. Trading in these markets occurs based on traditional price setting mechanisms or more realistic mechanisms inspired by electronic markets.

Definition of the Subject

Finance can be broadly defined as studying the allocation of resources over time in an uncertain environment. Consumers are interested in saving part of their current income and transfer it for consumption in the future (e. g., saving for retirement). On the other hand, firms are looking to raise capital to finance productive investments that will payoff in the future. In both decisions, the future is uncertain and individuals and firms are required to evaluate the risks involved in buying an asset (e. g., stocks and bonds) or investing in a project.

The traditional modeling approach in finance is to introduce strong assumptions on the behavior of agents. They are assumed to have perfect knowledge of the structure of the economy and to correctly process the available information. Based on these two assumptions, agents are able to form Rational Expectations (RE) such that their beliefs are not systematically wrong (in other words, the forecasting errors are random). Common sense suggests that these assumptions impose unreasonable requirements on the cognitive and computational abilities of agents. In practice, investors and firms are trying to learn to behave "*rationally*" in an economic system that is continuously evolving and where information is imperfect. In addition, there is an increasing amount of empirical evidence that is not consistent with RE theories.

These limitations have motivated an interest in finance to relax the strong assumptions on agents' behavior. Agent-based modeling contributes to this literature by assuming that consumers and firms have limited computational abilities (also known as bounded rationality) and learning (rather than knowing) the mechanisms governing the economy. These models have two main targets. First, to determine the conditions that lead a population of bounded-rational interacting agents to produce an aggregate behavior that resembles the one of a RE representative agent model. Second, they aim at explaining the empirical facts and anomalies that the standard approach fails to explain.

This entry is structured as follows. In Sect. "Introduction" we discuss in more detail the application of agent-based modeling in finance. In particular, most of the early literature has focused on one specific aspect of financial economics, asset pricing. Sects. "The Standard RE Model" to "Computational Agent-Based Models" introduce the standard asset pricing model and describe the agent-based approaches that have been proposed in the literature. Sect. "Other Applications in Finance" presents some (more recent) applications of agent-based models to corporate finance and market microstructure and, finally, Sect. "Future Directions" discusses some possible future directions on the application of agent-based models in finance.

Introduction

The goal of asset pricing models is to provide an explanation for the "fair" valuation of a financial asset paying an uncertain cash flow. A key role in asset pricing models is played by agents expectations regarding the future cash flow of the asset. Standard economic models assume that agents have *Rational Expectations* (RE). The RE hypothesis is the outcome of some more basic assumptions on agents behavior: they know and use all the information available, they have unlimited computational ability, and rationality is common knowledge in the population. Common sense and introspection suggest that these are quite strong assumptions if the target is to build a realistic model of agents behavior. A justification for assuming RE in asset pricing models is provided by [37]:

> … this hypothesis (like utility maximization) is not "behavioral": it does not describe the way agents think about their environment, how they learn, process information, and so forth. It is rather a property likely to be (approximately) possessed by the outcome of this unspecified process of learning and adapting.

Agent-based models try to address the issues left unspecified by the RE proponents: how do agents learn and process the information available? In other words, how do they form expectations? In fact, in the intent of the RE proponents, rationality is simply a property of the outcome (e. g., asset price) rather than an assumption about the subjective expectation formation process.

The innovative aspect of the agent-based approach is that it explicitly models "*this unspecified process of learning and adaptation*" (in Lucas's words). The common elements of the wide range of agent-based asset pricing models are:

Expectations agents hold subjective expectations that are *bounded rational*, that is, they are based on processing the available (and possibly imperfect and costly) information and that evolve over time. Agent-based models explicitly specify the way individuals form their expectation, instead of leaving it totally unspecified as in the RE approach.

Heterogeneity agents have different subjective expectations about the future due to heterogeneity in the way they process or interpret information. The RE setup suppresses agents heterogeneity: given the same information set, there is only one way to be rational and agents are thus homogeneous.

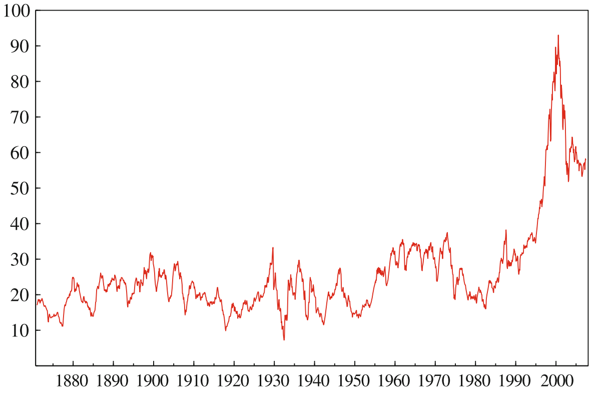
Evolution agents evolve in the sense that they abandon a belief if it performs poorly. Instead, rational models typically rely on the latent assumption that non-rational agents will not survive a (unspecified) process of evolutionary market competition.

Based on these basic ingredients, the agent based literature has now grown in different directions and we can distinguish two clearly defined approaches to agent-based modeling. The main difference between them is how they combine the different characteristics discussed above:

Analytical models these models assume that there are many expectation types and agents switch between them according to a deterministic or stochastic process. In the deterministic case, evolution is based on the past performance of the beliefs: agents discard belief types that perform badly compared to the other available. Instead, models with stochastic switching assume that a process governs the imitation and mutation of types, with possible additional features of herding. These models are simple and analytically tractable.

Computational models agents beliefs can change (or mutate) over time, due to the evolutionary selection of the best performing beliefs. Contrary to the analytical approach, the computational models renounce to analytical tractability in order to investigate more realistic expectation formation processes. Most of these models adopt fitness criteria (e. g., a Genetic Algorithm) to model the evolution of expectations.

The first aim of the agent-based literature is to understand whether introducing less restrictive assumptions on agents behavior (bounded rationality, heterogeneity, and evolutionary selection of expectations) is consistent with the economy converging to the RE equilibrium. If this is the case, it can be argued that relaxing the homogeneity and rationality of agents represents a feasible way to describe the way individuals learn and adapt to achieve an outcome consistent with RE. The second aim of this literature is to provide an explanation for the empirical behavior of asset prices. To illustrate the main stylized facts of financial returns, we consider the Standard & Poors 500 (S&P500) Composite Index (a U.S. equity index). Figure 1 shows the Price-to-Dividend (PD) ratio from 1871 until 2006 at the monthly frequency. It is clear that the PD ra-



Finance, Agent Based Modeling in, Figure 1
Monthly Price-to-Dividend Ratio for the S&P500 Index from 1871 to 2006

tio fluctuates significantly with some periods of extreme valuations, as in the late 1990s. The debate on the reasons for these fluctuations has not reached (yet) a widely accepted conclusion. On the one hand, there are RE models that explain the variation in the PD ratio by changes in the risk premium, i. e., the ex-ante rate of return required by agents to invest in the risky asset. Instead, other models attribute these swings to irrational expectations of investors, that are prone to optimism (and overvaluation) when asset prices are increasing. The two explanations are not mutually excluding since both factors might contribute to explain the observed fluctuations of the PD ratio.

Figure 2 considers the S&P500 Index from 1977 until 2007 at the daily frequency. The figure shows the returns (defined as the percentage change of the price of a financial asset) and the absolute value of the returns. Figure 3 describes the statistical properties of returns, such as the histogram and the autocorrelation function of the returns and absolute returns. The main stylized facts of daily returns are:

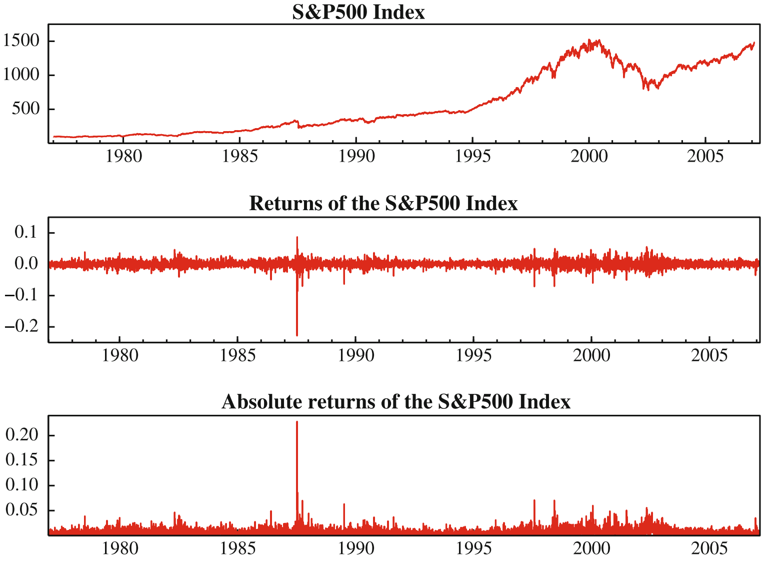
Volatility clustering returns alternate periods of high and low volatility (or variability). In calm periods, returns oscillate within a small range, while in turbulent periods they display a much wider range of variation. This is a feature common to different asset classes (e. g., equities, exchange rates, and bonds). The time series of returns and absolute returns on the S&P500 in Fig. 2 clearly show this pattern. In certain periods returns vary in a narrow interval between ±1%, while in other periods their variability is higher (e. g., between ±3 and 5%).

Leptokurtic distribution the distribution of asset returns has a sharp peak around the mean and fat tails (compared to the normal distribution). Large events (positive and negative) are more likely to occur compared to what is expected under the assumption of normality. This property emerges clearly from the top plot of Fig. 3 that shows the histogram of the S&P500 returns and the normal distribution (based on the estimated mean and variance).

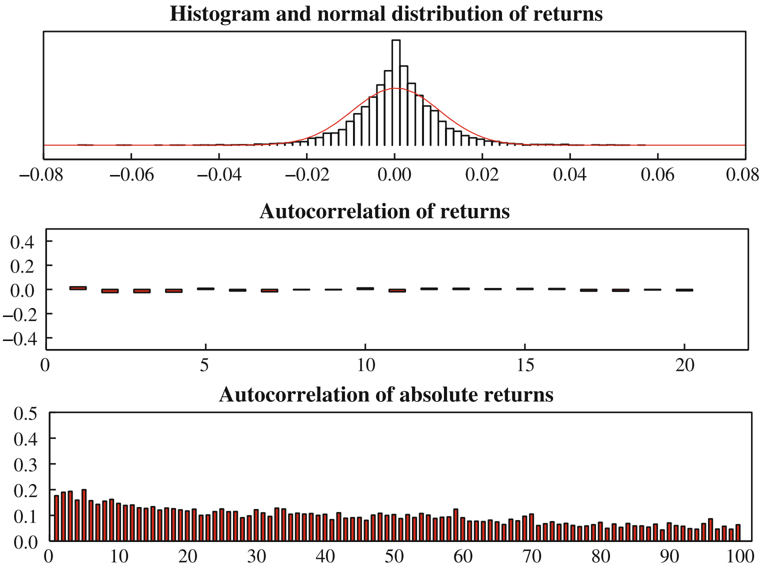
No serial correlation returns do not display significant linear serial correlation. The autocorrelation function of the returns (mid-plot of Fig. 3) is close to 0 at all lags considered.

Persistence in volatility on the other hand, volatility (measured by absolute or square returns) has significant linear dependence. The autocorrelation of the absolute returns in Fig. 3 is about 0.1 (and still significant) at lag 100.

Another relevant fact that is short of explanations is the large trading volume that occurs in financial markets. A model in which everybody is rational and knows that everybody else is rational cannot account for the existence of such relevant volume of trade. Agent-based models aim at explaining this phenomenon based on the assumption that agents hold heterogeneous expectations. Volume can



Finance, Agent Based Modeling in, Figure 2
Daily observations of the S&P500 Index from 1977 to 2007. (*top*) Time series of the Index, (*middle*) the returns, (*bottom*) the absolute value of returns



Finance, Agent Based Modeling in, Figure 3
Statistical properties of the S&P500 returns: (*top*) histogram and normal distribution, (*middle*) autocorrelation function (max lag 20) for the returns, (*bottom*) autocorrelation function (max lag 100) for the absolute returns

arise, for example, if an optimistic agent is willing to buy an asset from a pessimistic agent (that is willing to sell). An interesting feature of trading volume is its asymmetry during markets cycles: it is typically high when financial markets are booming, and low when the prices are decreasing. There is also empirical evidence that trading volume and volatility are correlated, suggesting that the same economic mechanism might be able to explain both phenomena.

Summarizing, the aim of the agent-based approach to asset pricing is to introduce more realistic assumptions on the way agents form expectations, learn from new information, and adapt to a changing environment. The research questions the agent-based approach is trying to answer are:

1. Under what conditions are these models able to reproduce the RE equilibrium (although starting from a more general setup where agents are not – a priori – assumed to have RE)?
2. Another issue is the empirical validity of these models: are they able to explain the empirical features of financial returns that standard RE models fail to account for?

In the following Sections, we describe some of the most well-known examples of agent-based models in finance, both in the *analytical* and *computational* group. However, we first introduce a basic RE model that is the starting point for most of the literature. In Sect. "Other Applications in Finance" we discuss other interesting applications of agent-based models in finance.

The Standard RE Model

We briefly consider the standard asset pricing model that is used as a benchmark in the agent-based literature. A more detailed discussion can be found in [25]. The model assumes that agent i faces the choice of investing her wealth among two assets: a riskless asset that pays a constant return r, and a risky asset that pays a stochastic dividend in period t denoted by D_t. A typical assumption is that agents have Constant Absolute Risk Aversion (CARA) preferences defined as $U(W_i) = -e^{-\lambda W_i}$, where $U(\cdot)$ indicates the utility function, W_i denotes the wealth of agent i and λ is the coefficient of absolute risk aversion. These preferences imply the following demand of shares of the risky asset, $X_{i,t}$:

$$X_{i,t} = \frac{E_{i,t}(P_{t+1} + D_{t+1}) - (1+r)P_t}{\lambda \sigma^2_{i,t}(P_{t+1} + D_{t+1})} , \qquad (1)$$

where P_t is the price of the risky asset in period t, $E_{i,t}(\cdot)$ is the conditional expectation of agent i about next-period payoff of the risky investment, and $\sigma^2_{i,t}(\cdot)$ is the conditional variance of the payoff for agent i. Agents buy shares of the risky asset ($X_{i,t} > 0$) if they expect the return of a share to be higher compared to investing the same amount (P_t) in the riskless asset.

The equilibrium price of the risky asset is such that the aggregate demand and supply are equal. Assuming that there are S number of shares of the risky asset available, the equilibrium condition is

$$S = \sum_i X_{i,t} \,. \qquad (2)$$

The aggregation across different individuals is simplified by assuming a representative agent with expectation $E_t(P_{t+1} + D_{t+1})$ (and similarly for the conditional variance) for all i's in the economy. This is equivalent to assume that agents are homogeneous in their expectation about the future payoff of the risky asset. In addition, assuming that the representative agent holds RE, it can be shown that the equilibrium price of the risky asset is a linear function of D_t given by

$$P_t = a + bD_t \,,$$

where a and b are constant (and related to the structural parameters of the model).

There is an extensive literature that aims at relaxing the strong restrictions imposed by the RE hypothesis. Models of *rational learning* assume that agents (typically in a representative agent setup) have to learn (rather than know) the structure of the economy, e. g., the parameters governing the cash flow process. In this case, agents are rational in the sense that they process optimally the information available. However, they do not hold rational expectations since they have imperfect knowledge of the structure of the economy. An alternative route followed in the literature is to consider the effect of *behavioral* biases in the expectation formation process. A comparison of the vast literature on rational learning and behavioral models is provided by [6].

Agent-based models build on these extensions of the basic asset pricing model by considering both rational learning and irrational expectations in a richer economic structure where agents hold heterogeneous expectations. We will now discuss some of the most well-known models in the analytical and computational agent-based literature and deal with their main differences.

Analytical Agent-Based Models

The analytical models assume that the population of agents can choose among a small number of beliefs (or predictors) about next period payoff of the risky asset. Heterogeneity is introduced by allowing agents with different predictors to co-exist, and learning might occur if they are allowed to switch between different beliefs in an evolutionary way.

These models can be described as follows. Assume there are a set of H belief types publicly available to agents. Denote the belief of type h (for $h = 1, \ldots, H$) about next period payoff by $E_{h,t}(P_{t+1} + D_{t+1})$ and the conditional variance by $\sigma^2_{h,t}(P_{t+1} + D_{t+1})$. Since these models depart from the assumption of RE, they typically introduce a behavioral assumption that the beliefs are either of the *fundamentalist* or the *trend-following* type. [20] and [21] conducted survey studies of exchange rate traders and found that their expectations could be represented as trend-following in the short-run, but fundamentalist in the long run. Fundamentalist expectations are characterized by the belief that the market price is anchored to the asset fundamental valuation and deviations (of the price from the fundamental) are expected to disappear over time. In this case, the belief uses both information about the asset price and the dividend process (that drives the fundamental value) to form an expectation about the future. On the other hand, trend-following expectations exploit only information contained in the price series to extrapolate the future dynamics. These types of beliefs are obviously not consistent with the RE equilibrium although they are supported by empirical evidence of their widespread use in financial markets.

Another key assumption of agent-based models concerns the evolution of beliefs: agents switch between expectations based on their past performance or because of interaction with other agents in the population. It is possible to distinguish two families of models with different evolutionary dynamics:

Deterministic evolution agents switch between the different beliefs based on a deterministic function. Typically, the switching is determined by past forecast accuracy of the predictors or their realized profits.

Stochastic evolution a stochastic process governs the switching of agents between beliefs.

Deterministic Evolution

An example of an agent-based model with deterministic evolution is proposed by [7]. A simple version of their model assumes that there are only two types of beliefs: fundamentalists and trend-followers. Some simplifying assumptions are used in deriving the equilibrium price: the dividend process D_t is assumed to be $i.i.d$ (with mean $\bar{D}$) and agents have homogeneous expectations about the dividend process. In this case, the expectation about next period payoff $E_{h,t}(P_{t+1} + D_{t+1})$ in Eq. (1) becomes $E_{h,t}(P_{t+1}) + \bar{D}$.

Lets denote by $P^*(= \bar{D}/r)$ the constant RE fundamental price. The fundamentalist type has the following belief:

$$E_{\mathrm{F},t}(P_{t+1}) = P^* + g_\mathrm{F}(P_{t-1} - P^*) . \tag{3}$$

When $0 < g_\mathrm{F} < 1$, fundamentalists believe the asset price will revert toward its fundamental value, and g_F can be interpreted as the speed at which this adjustment is expected to occur. This model assumes that when agents form their belief at time t they actually do not observe the realized asset price for period t. This explains the fact that the expectation is a function of the last observed price, P_{t-1}.

Brock and Hommes assume that agents pay a cost C to acquire the fundamentalist predictor. The underlying idea is to let them choose whether to buy a "sophisticated" predictor (that requires calculating the fundamental value) or, alternatively, to extrapolate from past realized prices. The belief of the trend-followers is given by:

$$E_{\mathrm{TF},t}(P_{t+1}) = g_\mathrm{TF} P_{t-1} . \tag{4}$$

The value of the parameter g_TF determines the strength of extrapolation of the trend-followers. If $g_\mathrm{TF} > 1$, they expect an upward trend in prices and, vice versa, for $0 < g_\mathrm{TF} < 1$.

Assuming the supply of the risky asset, S, in Eq. (2) is equal to 0, the equilibrium asset price, P_t, is given by:

$$P_t = \left(\frac{n_{\mathrm{F},t}(1 - g_\mathrm{F}) - r}{1 + r} \right) P^* + \left(\frac{n_{\mathrm{F},t}(g_\mathrm{F} - g_\mathrm{TF}) + g_\mathrm{TF}}{1 + r} \right) P_{t-1} , \tag{5}$$

where $n_{\mathrm{F},t}$ indicates the fraction of agents in the population using the fundamentalist belief and the remaining $n_{\mathrm{TF},t}(= 1 - n_{\mathrm{F},t})$ using the trend-following one. [7] assumes the evolution of the fractions $n_{\mathrm{F},t}$ is governed by a discrete choice probability model:

$$n_{\mathrm{F},t} = \frac{1}{1 - \exp\left[\beta(U_{\mathrm{TF},t-1} - U_{\mathrm{F},t-1})\right]} , \tag{6}$$

where $U_{h,t-1}(h = F, TF)$ is a measure of the fitness of belief h defined as:

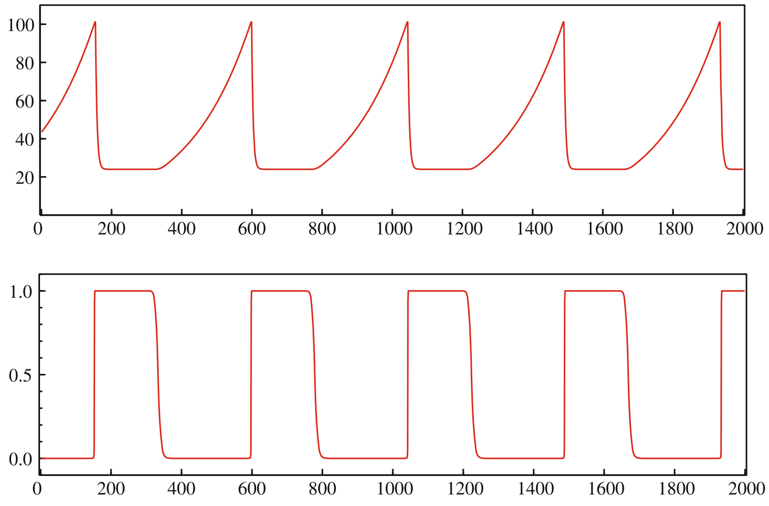
$$U_{\mathrm{F},t-1} = \pi_{\mathrm{F},t-1} + \eta U_{\mathrm{F},t-2} - C, \quad \text{and}$$
$$U_{\mathrm{TF},t-1} = \pi_{\mathrm{TF},t-1} + \eta U_{\mathrm{TF},t-2} ,$$

where $\pi_{h,t-1}$ measures the fitness performance (measured by realized profits or forecast accuracy) of the belief h at time $t - 1$ and η is a parameter that determines the memory in the performance measure. C in $U_{\mathrm{F},t-1}$ represents the cost that agents face if they adopt the fundamentalist belief (while the trend-following is available at no cost). The fraction in Eq. (6) depends on the parameter $\beta(> 0)$ that determines the speed at which agents respond to differentials of performance among beliefs. If β is small, agents are very reluctant to switch and require a significantly large difference in fitness to adopt another predictor. On the other hand, when β is large, even small differences of performance cause dramatic changes in the fractions. For a given value of β, if the fundamentalist belief significantly outperforms the trend-following (that is, $U_{\mathrm{F},t-1} \gg U_{\mathrm{TF},t-1}$), then the fraction $n_{\mathrm{F},t-1}$ tends to 1, meaning that most agents in the economy switch to the fundamentalist expectation.

The interesting feature of this model is that it can converge to the RE equilibrium or generate complicated dynamics depending on the value of the parameters. For some combinations of the g_F and g_TF parameters, the system converges to the RE equilibrium (i. e., the deviation is equal to 0). However, trend-followers can destabilize the economy when their extrapolation rate, g_TF is high enough. For small values of β the dynamical system converges to the RE equilibrium. However, for increasing values of β the system experiences a transition toward a non-fundamental steady state and complicated dynamics (limit cycles and strange attractors) emerge.

In the presence of information cost (to buy the fundamentalist predictor) and evolutionary switching between strategies, the economy might still converge to the RE equilibrium for a large set of parameter values. However, it is also able to generate large fluctuations of the asset price around the fundamental value. Figure 4 shows a time series of the asset price P_t and the fraction of fundamentalists described in Eqs. (5) and (6). The constant fundamental value in this Figure is equal to 25. As it is clear from the picture, the asset price experiences large swings away from the fundamentals that are explained by the increased importance of agents using the trend-following belief. When the mispricing becomes too large, the economy experiences a sudden change of sentiment with most agents switching to the fundamentalist belief. In this sense, the model is more appropriate to explain the boom-bust dynamics of financial markets.

Although the purely deterministic model captures the relevant features of the dynamics of financial markets, adding a stochastic component provides simulated series that better resemble the observed ones (such as Fig. 1). The model can be extended by considering an approximation error term in Eq. (5) that interacts with the dynamics of the model. Figure 5 shows the asset price and the fraction of fundamentalists for a normally distributed error term with standard deviation equal to 2. The asset price shows large and persistent deviations from the fundamental value



Finance, Agent Based Modeling in, Figure 4

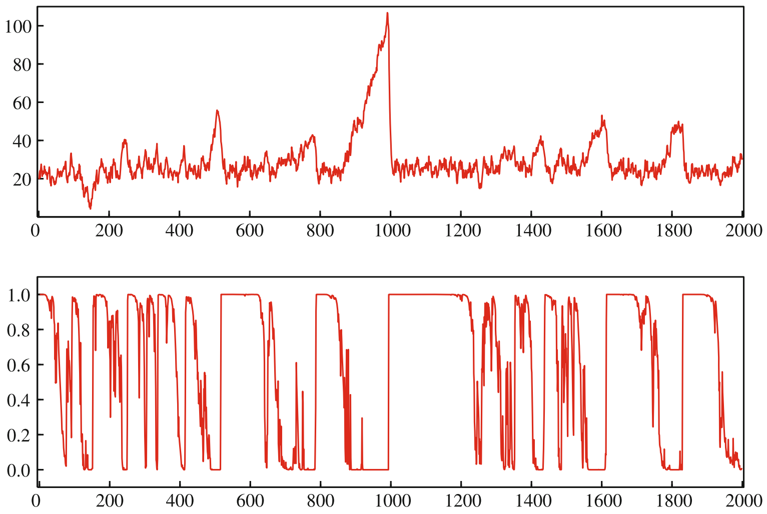
Brock and Hommes model with 2 belief types, fundamentalists and trend-followers. The top plot represents a time series of the asset price and the bottom plot depicts the fraction of fundamentalists, $n_{F,t}$. The parameters of the model: intensity of choice $\beta = 0.5$, the interest rate $r = 0.0083$, the parameter of the fundamentalists $g_F = 0.8$, the parameter of the trend-followers belief $g = 1.014$, the cost of the fundamentalist predictor $C = 0.5$, memory parameter $\eta = 0.99$

($P^* = 25$), in some periods as extreme as reaching 100 while, in other periods, more moderate. Since the dividend process is assumed to be *i.i.d.*, the price can also be interpreted as a valuation (PD) ratio. Comparing the properties of this time series with the one for the S&P500 in Fig. 1, it seems that it is able to capture its main qualitative features. [5] provide empirical evidence of the ability of a similar model to explain the long-run behavior of stock prices.

The model proposed by Brock and Hommes is an example of a literature interested in the (possible) emergence of complicated dynamics in asset pricing models. An early contribution to the deterministic literature is [15]. They assume that there are two types of investors: some extrapolating from past deviations while the other group is more sophisticated and able to evaluate whether the asset is over- or under-valued (and sells or buys more aggressively if the mispricing is large). A third actor in the model is the market maker. The role of the market maker is to aggregate demand and supply and to fix the price of the asset. This mechanism is different from the assumption in Eq. (2). In that case, agents submit their demand function (quantity as a function of price) and the price is set at the value that clears the market. Instead, the market maker receives orders from the agents and moves the price to offset excess demand or supply. This represents a disequilibrium mechanism since market makers use their inventory of stocks to provide liquidity in case of excess demand and accumulate stocks in case of excess supply. The results for this model are similar to what was discussed above. The RE equilibrium is obtained when the sophisticated agents dominate the market. However, limit cycles and chaos arise when the trend-following agents are relatively important and the economy fluctuates between periods of increasing asset prices and sudden crashes. Another model that assumes the market maker mechanism is proposed by [9]. In this case, agents form their expectations based on either the fundamental or extrapolative approach. However, the excess demand function of the chartist is assumed to be nonlinear. When the extrapolation rate of the chartist is sufficiently high, the system becomes unstable and limit cycles arise. While these early models assumed that the fractions of agents are fixed, [16] and [17] introduced, in a similar setup, time-variation in those fractions. The driving force for the variation of the fractions is the relative performance of the beliefs (similar to what we discussed above for the model of Brock and Hommes). Some of the more recent models that extend these early contributions are [10,11,12,18,19,24,50], and [51]. A comprehensive survey of the literature is provided in [26].

Stochastic Evolution

An early example of an agent-based model in which individuals switch in a stochastic fashion was proposed by [27]. He uses a slightly different setup compared to the Standard RE Model. In his model the asset is a for-



Finance, Agent Based Modeling in, Figure 5
Same model and parameter values used in Fig. 4 with an error term added to Eq. (5) that is normally distributed with mean zero and standard deviation equal to 2

eign bond and the agent has to decide whether to invest at home (at the riskless interest rate r) or buy a unit of foreign currency and invest abroad (at the risky interest rate ρ_t, assumed to be normally distributed with mean ρ and variance σ^2_ρ). The price P_{t+1} represents the exchange rate. The only difference with the model described earlier is that in the demand of type h agent in Eq. (1), $E_{h,t}(P_{t+1} + D_{t+1})$ is replaced by $(1 + \rho)E_{h,t}(P_{t+1})$. The fundamental value of the asset in this model is assumed to evolve as a random walk, that is, $P^*_t = P^*_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2_\epsilon)$.

Similarly to the previous model, there are two types of beliefs: fundamentalists and chartists. The fundamentalist belief is the same as in Eq. (3), while the chartists have belief given by:

$$E_{\mathrm{TF},t}(P_{t+1}) = (1 - g_{\mathrm{TF}})P_t + g_{\mathrm{TF}}P_{t-1}\ .$$

The switching between beliefs in Kirman's model is driven by two mechanism: social interactions and herding. Interaction means that agents meet in pairs and communicate about their beliefs. The result of this communication is that, with a probability $(1 - \delta)$, an agent changes her belief to the one of the other agent. In this model, market information (such as prices or dividends) do not play any role in the decision of the agents to adopt the fundamentalist or trend-following beliefs. This is in sharp contrast to the model of [7] where the selection of the belief is endogenous and based on their past performance. In addition to the probability of switching belief due to social interaction, there is a probability ϵ that an agent independently changes belief. If we denote by $N_{\mathrm{F},t}$ the number of agents in the population (N is the total number of agents) using the fundamentalist belief at time t, Kirman models the evolution from $N_{\mathrm{F},t-1}$ to $N_{\mathrm{F},t}$ according to a markov chain with the following transition probabilities:

$$\begin{aligned}
P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 1) =& \left(1 - \frac{N_{\mathrm{F},t-1}}{N}\right) \\
& \left(\epsilon + (1-\delta)\frac{N_{\mathrm{F},t-1}}{N-1}\right) \\
P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = -1) =& \frac{N_{\mathrm{F},t-1}}{N} \\
& \left(\epsilon + (1-\delta)\frac{N - N_{\mathrm{F},t-1}}{N-1}\right) \\
P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 0) =& 1 - P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 1) \\
& - P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = -1)\ .
\end{aligned}$$

The second part of the opinion formation can be characterized as herding. Kirman assumes that the agents receive a noisy signal, $q_{i,t}$, about the fraction of the population that is fundamentalist:

$$q_{i,t} = \frac{N_{\mathrm{F},t}}{N} + \xi_{i,t}\ ,$$

where $\xi_{i,t}$ is distributed as $N(0, \sigma^2_\xi)$ and $i = 1, \ldots, N$. Based on this signal about the average opinion in the economy, agents herd by coordinating in adopting the belief that is more popular. Agent i uses the fundamentalist belief if her signal, $q_{i,t}$, is larger than 0.5 and the trend-following belief otherwise. The fraction of agents using the fundamentalist belief (denoted by $n_{\mathrm{F},t}$) is then given by:

$$n_{\mathrm{F},t} = \frac{1}{N} \sum_{i=1}^{N} I(q_{i,t} \geq 0.5) \, .$$

Given the beliefs of the two types, the fractions and assuming that the supply of foreign currency is proportional to the time varying fundamental value, the equilibrium price of the model is given by:

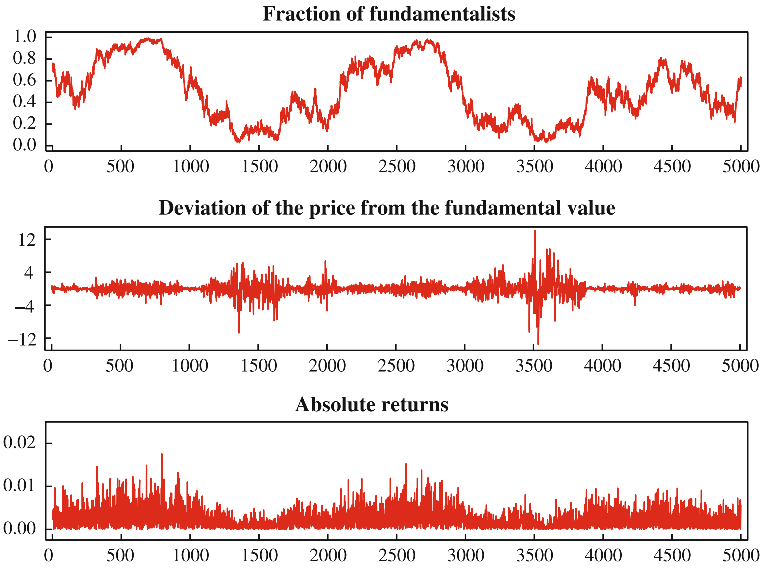
$$P_t = \frac{n_{\mathrm{F},t} - \gamma}{A} P_t^* - \frac{n_{\mathrm{F},t} g_{\mathrm{F}}}{A} P_{t-1}^* + \frac{(1 - n_{\mathrm{F},t}) g_{\mathrm{TF}}}{A} P_{t-1} \, , \quad (7)$$

where the constants γ and A are functions of the structural parameters of the model.

Figure 6 shows a time series simulated from Kirman's model. The fraction of agents using the fundamentalist belief, $n_{\mathrm{F},t}$, displays significant variation over time, with some periods being close to 1 (most agents are fundamentalists) and other periods close to zero (trend-followers dominate). The resulting price dynamics can be characterized as follows. When fundamentalists dominate the market, the asset price tracks closely the fundamental value and returns volatility is high. On the other hand, when trend-followers dominate the market the price tends to deviate significantly from the fundamental and volatility is lower. The time series provide a clear intuition about the ability of the model to account for periods of large deviation from the fundamentals and of persistent volatility.

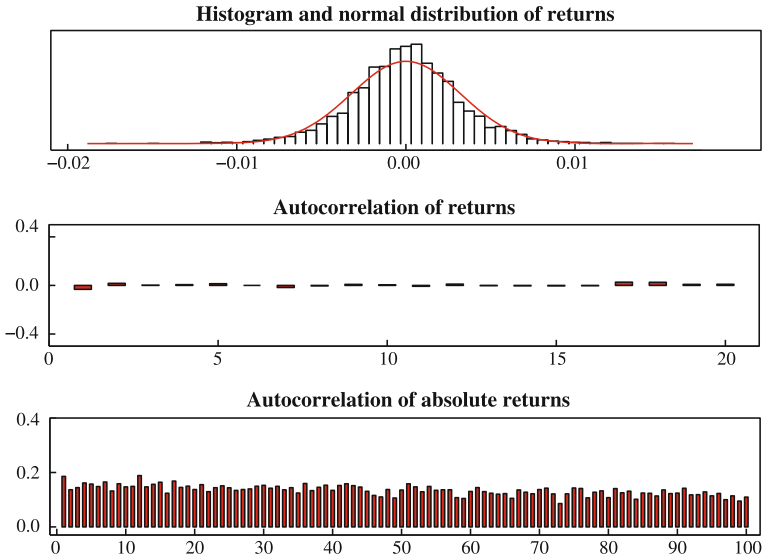
A main objective of this model is to provide an explanation for the stylized facts of financial returns that were discussed earlier. Figure 7 shows some of the statistical properties for the simulated series. The histogram shows the typical leptokurtic property of financial returns. The distribution of the simulated returns shows a higher concentration of probability mass around zero and in the tails (compared to the normal distribution). The returns autocorrelations are very close to zero and statistically insignificant. However, the absolute returns show significantly positive and slowly-decaying autocorrelations. Hence, the simulated series from Kirman's model are able to replicate the relevant empirical features of daily financial returns.

[41] and [42] propose a model inspired by the opinion formation mechanism of Kirman. The model assumes that agents are either fundamentalists or chartists. In addition, the chartist group is composed of agents that are either optimistic or pessimistic. Agents can switch between the two sub-groups due to herding (following the majority opinion) and also to incorporate the recent trend in asset prices. Instead, the switching between fundamentalist and chartist beliefs is based on the excess profits of the rules. In this aspect, the model allows for a feedback effect from market price to the fractions similarly to [7].



Finance, Agent Based Modeling in, Figure 6

Time series generated from the model proposed by [27] for the following parameter values: $N = 1000$, variance of the $\sigma_\epsilon^2 = 10$, $\rho = 0.00018538$, $r = 0.000133668$, $g_{\mathrm{F}} = 0.6$, $g_{\mathrm{TF}} = 0.475$, $\delta = 0.10$, $\epsilon = 0.000325$, and $\sigma_\xi^2 = 0.43/N$. The top plot shows the fraction of fundamentalist agents in the population, the middle plot the deviation of the market price from the fundamental value, and the bottom plot displays the absolute value of the asset returns



Finance, Agent Based Modeling in, Figure 7

Statistical properties of the time series in Fig. 6. The top plot shows histogram of the series and the parametric distribution under the assumption of normality, and the bottom plots show the autocorrelation of the returns and absolute returns, respectively

A market maker mechanism aggregates the demand for the risky asset of the agents and determines the prices. Another model based on interacting agents and herding behavior is proposed by [14]. They model the communication among (groups of) agents as a random graph and the returns dynamics closely match the stylized facts of financial returns. [26] and [49] provide extensive overviews of models based on random communication in a population of interacting agents.

Computational Agent-Based Models

Computational agent-based models move one step further compared to analytical models. The setup is very similar: the simple asset pricing model described above, the assumption of heterogeneity of beliefs in the population, and evolutionary pressure to use the best performing predictors. However, computational models do not assume a priori the form of agents' beliefs. Instead, they let agents learn, adapt and explore a large set of strategies and use the best performing ones. In this sense these models allow to investigate the *emergence* of trading strategies and their survival in the market. A relevant question, and an unsettled dispute between academics and practitioners, is the role and importance of technical analysis. Computational agent-based models do not assume (a priori) that trend-following rules are used by agents (as in the analytical approach), but allow for these rules to emerge from the evolutionary and learning processes. Hence, they can indicate the conditions that lead to the emerge and survival of trend-following rules in the market.

One of the first and most famous example of a computational agent-based model is the Santa Fe Institute (SFI) artificial market proposed by [3]. As mentioned above, the key feature of this and similar models is the way the expectation formation process is represented.

Each agent in the economy is endowed with a set of predictors, in the form of *condition/forecast* rules. These rules are a form of classifier system that identify a state of the world and indicate an action (in this case a forecast of future returns). Each agent in the economy is assumed to rely on a set J of market predictors (classifier rules) that consist of two elements:

Condition a set of bit-strings that characterize different possible states of the market. Each bit represents a state of the world, and the design of the SFI market allows for a set of bits related to *fundamentals* (that relate the asset price to the underlying dividend process) and another set of *technical* bits (that relate the current price to a moving-average of past prices of different length). The role of the bit-strings is to provide the agent with the ability to identify the current state of the market.

Forecast associated with each bit-string j (for $j = 1, \dots, J$) is a parameter vector (a_j, b_j) that together with the linear forecasting rule $E^j_{i,t}(P_{t+1}+D_{t+1}) = a_j(P_t + D_t) + b_j$ provides agent *i* with the forecast for next period payoff. The agent then combines the forecast of the *H* most accurate predictors that are active, based on the observed market condition.

The next building block of the SFI artificial market is the learning process. This is implemented using a Genetic Algorithm (GA) that enables learning in both the condition and the forecast part of the classifier. Agents have a probability *p* to update their predictors through learning in every period. The frequency of learning (measured by *p*) plays a relevant role in the resulting dynamics of the model since it identifies how quickly agents adapt to changes in the environment and respond to it. In the learning phase, 15% of the worst performing predictors are dropped, and new rules are generated using a GA with uniform crossover and mutation.

The aim of the model is to test the hypotheses discussed above:

1. Does the SFI market converge to the RE equilibrium?
2. Can the SFI market explain the stylized facts of financial returns?

It turns out that the answer is positive to both questions, depending on the speed of learning parameter *p*. This parameter plays a crucial role in the resulting dynamics of the SFI market and two regimes can be identified:

Slow-learning in this case the agents are engaged in learning (via the GA) every 1000 periods (on average). The resulting price dynamics shows convergence to the RE equilibrium characterized by agents having homogeneous expectations, negligible trading volume (although some occurs when agents change their beliefs due to learning), and returns that are normal and homoskedastic. What is remarkable is that the convergence to the RE equilibrium is not built-in the model, but it is achieved by the learning and evolutionary process taking place in the SFI market. Another interesting result is that the technical trading bits of the classifier play no role and are never active.

Fast-learning in this experiment the agents update their predictors via learning (on average) every 250 periods. The price dynamics shows the typical features of financial time series, such as alternating periods of high and low volatility, fat tailed distribution, high trading volume, and bubbles and crashes. An interesting result of the fast-learning experiment is the *emergence* of the trend-following rules. The technical trading bits of the predictors are activated and their effect on the asset price spurs even more agents to activate them. In this sense, trend-following beliefs emerge endogenously in the economy and they are not eliminated in the evolution of the system, but survive. This is a quite relevant result also from an empirical point of view. As we mentioned earlier, technical analysis is widely used by practitioners and the SFI market provides an explanation for its emergence (and survival).

[29,30], and [31] have recently proposed an artificial market that is largely inspired by the SFI market. However, LeBaron changed some very relevant assumptions compared to the earlier model. An innovation in this new artificial market is the assumption on the preferences of agents. While the SFI market (and many analytical models) rely on CARA preferences, [29] considers Constant Relative Risk Aversion (CRRA) preferences. In this case, wealth plays a role in the demand of agents (while with CARA does not) and, consequently, allows for differential market impact of agents based on their wealth. Another innovation concerns the expectation formation process. [29] departs from the SFI market assumption of different "*speed of learning*" across agents. Instead, LeBaron assumes that agents have different memory length in evaluating strategies. In every period agents assess the profitability of the strategy adopted. However, agents evaluate their strategy using different backtesting periods: some agents test their strategies on only the last few months, while other agents consider the last 20 years. In this sense, they are heterogeneous in their *memory* rather than in the speed of learning. Another feature of this model is that the classifier predictor is replaced by a neural network. The learning and evolution is always characterized by a GA. Despite these innovations, the earlier results of the SFI market are confirmed: a market populated by long-memory agents converges to the RE equilibrium. However, in an economy with agents holding different memory lengths, the asset price series shows the typical features of financial returns (no serial correlation, volatility clustering, fat-tailed distribution, high trading volume, and correlation between volume and volatility).

Another recent artificial stock market model is proposed by [8]. The setup is the simple asset pricing model described in Sect. "The Standard RE Model". Chen and Yeh assume that the expectation of agent *i* about next period payoff is of the form $E_{i,t}(P_{t+1}+D_{t+1}) = (P_t+\mu)(1+\theta_1 \tanh(\theta_2 f_{i,t}))$. The quantity $f_{i,t}$ characterizes the expectation of the agent and it evolves according to genetic programming. If $f_{i,t}$ is equal to zero the agent believes in the efficiency and rationality of the market, that is, expects the asset price tomorrow to increase by the expected dividend growth rate.

Compared to the SFI market, they adopt a Genetic Programming (GP) approach to model agents' learning and evolution. The model assumes that agents evolve due to two types of pressures: peer-pressure (the agent performance compared to the rest of the population) and self-pressure (own evaluation of the performance). The probability that an agent searches for better forecasting rules depends on both forms of pressure. If agents rank low in terms of performance compared to their peers, then the probability that they will search for other forecasting rules is higher. The population of forecasting rules evolves due to competition with new rules that are generated by applying genetic operators (reproduction, cross-over, mutation) to the existing rules. The rules space evolves independently of the rules adopted by the agents. When an agent decides to search (due to the pressures mentioned above), forecasting rules are randomly selected from the population until a rule is found that outperforms the one currently adopted by the agent. Chen and Yeh show that the price dynamics of the model is consistent with an efficient market. The investigation of the statistical properties of the returns generated by the model shows that the series does not have any linear and nonlinear dependence, although there is some evidence for volatility clustering. Analyzing the type of rules that agents use, they show that only a small fraction of them are actually using forecasting rules that are consistent with an efficient market (in the sense that they believe that $E_{i,t}(P_{t+1} + D_{t+1}) = P_t + \mu$ in which case $f_{i,t}$ is equal to 0). In other words, although a large majority of agents uses rules that imply some form of predictability in asset returns, the aggregation of their beliefs delivers an asset price that looks "*unpredictable*". In this sense they claim that the efficiency (or unpredictability) of the artificial market is an emerging property that results from the aggregation of heterogeneous beliefs in the economy. Another property that emerges from the analysis of this market is the rationality of a representative agent. Chen and Yeh consider the expectation of a representative agent by averaging the expectations across the agents in the economy. The forecasting errors of this "*representative*" expectation indicate that they satisfy a test for rationality: there is no evidence of systematic forecasting errors in the expectation (in statistical terms, the errors are independent).

Evolution and learning (via GA) have received quite a lot of attention in the literature. Other artificial-market models have been proposed in the literature. Some early examples are [4], and [46]. Some more recent examples are [1,2,47,48]. [32] is an extensive survey of the computational agent-based modeling approach.

Other Applications in Finance

The common theme across the models presented above is to show that departing from a representative rational agent is a viable way to explain the empirical behavior of asset prices. The more recent agent-based literature has shifted interest toward nesting this type of models in more realistic market structures. There are two typical assumptions used in the agent-based literature to determine the asset price: a market clearing or a market maker mechanism. Recent progress in the analysis of the micro-structure of financial markets has indicated the increasing importance of alternative trading mechanisms, such as order-driven markets. In these markets, traders decide whether to buy (sell) using a market or limit order. A market order means that the trader is ready to buy (sell) a certain quantity of stocks at the best available price; instead, with limit orders traders fix both a quantity of shares and a price at which they are willing to buy (sell). Limit orders are stored in the book until a matching order arrives to the market. They are different from quote-driven markets, where a market maker provides quotes and liquidity to investors. This has spurred a series of articles that propose agent-based models in this more realistic market structure. In particular, [13,36,44], and [45] consider an order-driven market where agents submit limit orders. Typically these models make simple behavioral assumptions on the belief formation process and do not consider learning and evolution of agents' expectations (typical of the computational agent-based models). In this sense, these models are closer to the stochastic agent-based approach reviewed in Sect. "Analytical Agent-Based Models". Recently, [33] has proposed a computational model for an order-driven market in which strategies evolve by imitation of the most successful rules.

[13] propose an order-driven market model in which the demand for the risky asset of the agents is determined by a fundamentalist, a chartist, and a noise component. The agents share the same demand function but the weights on the components are different across agents. Simulating the model suggests that the stylized facts of financial returns can be explained when all behavioral components (fundamentalist, chartist, and noise) participate to determine agents' beliefs. An additional feature of this setup is that it accounts for the persistence in the volatility of returns and trading volume. Such a micro-structural model allows also to investigate the effect of some key market design parameters (such as tick size, liquidity, and average life of an order) on the price formation process.

[44] consider a market structure where agents submit limit orders and the price is determined by market

clearing of supply (sell orders) and demand (buy orders) schedules. The behavioral assumptions are closely related to the clustering approach of [14]: a probabilistic mechanism governs the formation of clusters and, within a clusters, all agents coordinate in buying or selling the risky asset. Another behavioral assumption introduced in this model concerns the (positive) relation between market volatility and the limit order price. When the volatility is low, agents set the price of their limit order close to yesterday's asset price. However, when the market is experiencing wide swings in prices, agents' set limit prices that are significantly above or below yesterday's price for their orders. The results suggest that the model is able to explain the main stylized facts of financial returns. [45] consider an economy with a similar market structure but more sophisticated assumption on agents' behavior. They assume the population is composed of four types of agents: random traders (with 50% probability to buy or sell), momentum (buy/sell following an increase/decrease in prices), contrarian (act in the opposite direction of momentum traders), and fundamentalists (buy/sell if the price is below/above the fundamental value). They simulate the model in a way that non-random agents do not affect the asset price. The idea is to investigate the survival of these types of agents in the economy without affecting the aggregate properties of the model. They show that, on average, the fraction of wealth of momentum agents decreases while it increases for fundamentalist and contrarian traders.

Another recent paper that models an order-driven market is [36]. Agents can submit market and limit orders. They introduce the behavioral assumption that the agents are all fundamentalists, although they are heterogeneous in their belief of the fundamental value of the asset. They show that simulated series from this simple model follow a leptokurtic distribution and attribute this property to the structure of the market (rather than the behavioral assumptions). The same result is also obtained when random traders are considered. However, they are not able to explain other stylized fact such as the autocorrelation structure of volatility. This paper is interesting because it suggest that some of the stylized facts discussed earlier might not be related to agents' bounded rationality, but rather to the details of the market mechanism that is typically neglected in the agent-based literature.

Another area of application of agent-based modeling is corporate finance. [43] propose an agent-based model to investigate the security issuance preferences of firms and investors. The empirical evidence indicates that there is a clear dominance of debt financing, compared to other instruments, such as equities and convertible debt. This is a puzzle for theoretical models where it is customarily assumed that the payoff structure of the financing instruments are common knowledge. Under this assumption, the price should reflect the different characteristics of the securities and investors should be indifferent among them. Noe et al. consider a model that departs from the assumption of perfect knowledge about the security characteristics, and assume that firms and investors are learning (via a GA) about the profitability of the different alternatives. Debt and equity imply different degrees of risk-sharing between investors and firms: in particular, debt provides lower risk and return, contrary to equities that have a more volatile payoff structure. Investors' learning about the risk-return profile of the different instruments leads to the prevalence of debt on equity or convertible debt. Another conclusion of this model is that learning is imperfect: agents learn to price specific contracts and have difficulties in dealing with contracts that rarely occur in the market.

Future Directions

Agent-based modeling in finance has had a significant impact in shaping the way we understand the working of financial markets. By introducing realistic behavioral assumptions, agent-based models have demonstrated that they provide a coherent explanation for many empirical findings in finance. In addition, they are also able to provide a framework to explain how aggregate rationality can emerge in a population of bounded rational learning agents.

The strength of the agent-based approach is the ability to specify in greater detail the agents' behavior and the structure of market interactions. Simple agent-based models use realistic assumptions and can be solved analytically. However, they sacrifice the important aspect of the emergence of aggregate pattern based on agents' learning. This can be achieved by computational agent-based models. Since the approach is not bounded by the analytical tractability of the model, very detailed (and potentially more realistic) assumption can be introduced. However, this can represent a weakness of the approach since it might lead to over-parametrized models where it is hard to disentangle the role played by each of the assumptions on the aggregate behavior. In this sense, agent-based modeling should aim at balancing parsimony and realism of agents' description.

As already suggested in Sect. "Other Applications in Finance", the application of agent-based models is not limited to asset pricing issues. Recently, they have been used in corporate finance and market microstructure. This is certainly a trend that will increase in the future since these models are particularly suitable to investigate the interaction of market structure and agents' behavior.

Bibliography

Primary Literature

1. Arifovic J (1996) The Behavior of the Exchange Rate in the Genetic Algorithm and Experimental Economies. J Political Econ 104:510–541
2. Arifovic J, Gencay R (2000) Statistical properties of genetic learning in a model of exchange rate. J Econ Dyn Control 24:981–1006
3. Arthur WB, Holland JH, LeBaron B, Tayler P (1997) Asset pricing under endogeneous expectations in an artificial stock market. In: Lane D, Arthur WB, Durlauf S (ed) The Economy as an Evolving Complex System II. Addison–Wesley, Reading
4. Beltratti A, Margarita S (1992) Evolution of trading strategies among heterogeneous artificial economic agents. In: Wilson SW, Meyer JA, Roitblat HL (ed) From Animals to Animats II. MIT Press, Cambridge, pp 494–501
5. Boswijk HP, Hommes CH, Manzan S (2007) Behavioral heterogeneity in stock prices. J Econ Dyn Control 31:1938–1970
6. Brav A, Heaton JB (2002) Competing Theories of Financial Anomalies. Rev Financ Stud 15:575–606
7. Brock WA, Hommes CH (1998) Heterogeneous beliefs and and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
8. Chen S-H, Yeh C-H (2002) On the emergence properties of artificial stock markets: the efficient market hypothesis and the rational expectations hypothesis. J Econ Behav Organ 49:217–239
9. Chiarella C (1992) The dynamics of speculative behavior. Ann Oper Res 37:101–123
10. Chiarella C, He XZ (2001) Asset price and wealth dynamics under heterogeneous expectations. Quant Financ 1:509–526
11. Chiarella C, He XZ (2002) Heterogeneous Beliefs, Risk and Learning in a Simple Asset Pricing Model. Comput Econ 19:95–132
12. Chiarella C, He XZ (2003) Heterogeneous beliefs, risk, and learning in a simple asset-pricing model with a market maker. Macroecon Dyn 7:503–536
13. Chiarella C, Iori G (2002) A simulation analysis of the microstructure of double auction markets. Quant Financ 2: 346–353
14. Cont R, Bouchaud JP (2000) Herd behavior and aggregate fluctuations in financial markets. Macroecon Dyn 4:170–196
15. Day R, Huang W (1990) Bulls, bears and market sheep. J Econ Behav Organ 14:299–329
16. de Grauwe P, Dewachter H (1993) A chaotic model of the exchange rate: The role of fundamentalists and chartists. Open Econ Rev 4:351–379
18. de Grauwe P, Grimaldi M (2005) The exchange rate and its fundamentals in a complex world. Rev Int Econ 13:549–575
17. de Grauwe P, Dewachter H, Embrechts M (1993) Exchange Rate Theory – Chaotic Models of Foreign Exchange Markets. Blackwell, Oxford
19. Farmer JD, Joshi S (2002) The price dynamics of common trading strategies. J Econ Behav Organ 49:149–171
20. Frankel JA, Froot KA (1987) Using survey data to test standard propositions regarding exchange rate expectations. Am Econ Rev 77:133–153
21. Frankel JA, Froot KA (1990) Chartists, fundamentalists, and trading in the foreign exchange market. Am Econ Rev 80:181–185
22. Gaunersdorfer A (2000) Endogenous fluctuations in a simple asset pricing model with heterogeneous agents. J Econ Dyn Control 24:799–831
23. Gaunersdorfer A, Hommes CH (2006) A nonlinear structural model for volatility clustering. In: Kirman A, Teyssiere G (ed) Microeconomic models for long memory in economics. Springer, Berlin, pp 265–288
24. Goldbaum D (2005) Market efficiency and learning in an endogenously unstable environment. J Econ Dyn Control 29:953–978
25. Grossman SJ, Stiglitz JE (1980) On the Impossibility of Informationally Efficient Markets. Am Econ Rev 70:393–408
26. Hommes CH (2006) Heterogeneous agent models in economics and finance. In: Judd KL, Tesfatsion L (ed) Handbook of Computational Economics, vol 2: Agent-Based Computational Economics. North-Holland, Amsterdam, pp 1109–1185
27. Kirman A (1991) Epidemics of opinion and speculative bubbles in financial markets. In: Taylor MP (ed) Money and Financial Markets. Blackwell, Oxford, pp 354–368
28. Kirman A, Teyssiere G (2001) Microeconomics models for long-memory in the volatility of financial time series. Stud Nonlinear Dyn Econ 5:281–302
29. LeBaron B (2001) Empirical regularities from interacting long and short memory investors in an agent based in stock market. IEEE Trans Evol Comput 5:442–455
30. LeBaron B (2001) Evolution and time horizons in an agent based stock market. Macroecon Dyn 5:225–254
31. LeBaron B (2002) Short-memory traders and their impact on group learning in financial markets. In: Proceedings of the National Academy of Science: Colloquium, Washington DC, 99:7201–7206
32. LeBaron B (2006) Agent-based computational finance. In: Judd KL, Tesfatsion L (ed) Handbook of Computational Economics, vol 2: Agent-Based Computational Economics. North-Holland, Amsterdam, pp 1187–1233
33. LeBaron B, Yamamoto R (2007) Long-memory in an order-driven market. Physica A 383:85–89
34. Levy M, Levy H, Solomon S (1994) A microscopic model of the stock market. Econ Lett 45:103–111
35. Levy M, Solomon S, Levy H (2000) Microscopic Simulation of Financial Markets. Academic Press, New York
36. Licalzi M, Pellizzari P (2003) Fundamentalists clashing over the book: a study of order-driven stock markets. Quant Financ 3:470–480
37. Lucas RE (1978) Asset prices in an exchange economy. Econometrica 46:1429–1445
38. Lux T (1995) Herd Behaviour, Bubbles and Crashes. Econ J 105:881–896
39. Lux T (1997) Time variation of second moments from a noise trader/infection model. J Econ Dyn Control 22:1–38
40. Lux T (1998) The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions. J Econ Behav Organ 33:143–165

41. Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of interacting agents. Nature 397:498–500
42. Lux T, Marchesi M (2000) Volatility clustering in financial markets: A micro-simulation of interacting agents. Int J Theoret Appl Financ 3:675–702
43. Noe TH, Rebello MJ, Wang J (2003) Corporate financing: an artificial agent-based analysis. J Financ 58:943–973
44. Raberto M, Cincotti S, Focardi SM, Marchesi M (2001) Agent-based simulation of a financial market. Physica A: Stat Mech Appl 299:319–327
45. Raberto M, Cincotti S, Focardi SM, Marchesi M (2003) Traders'Long-Run Wealth in an Artificial Financial Market. Comput Econ 22:255–272
46. Rieck C (1994) Evolutionary simulations of asset trading strategies. In: Hillebrand E, Stender J (ed) Many-Agent Simulation and Artificial Life. IOS Press, Amsterdam
47. Routledge BR (1999) Adaptive Learning in Financial Markets. Rev Financ Stud 12:1165–1202
48. Routledge BR (2001) Genetic algorithm learning to choose and use information. Macroecon Dyn 5:303–325
49. Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Microscopic Models of Financial Markets. In: Schweitzer F (ed) Microscopic Models of Economic Dynamics. Springer, Berlin
50. Westerhoff FH (2003) Expectations driven distortions in the foreign exchange market. J Econ Behav Organ 51:389–412
51. Westerhoff FH (2004) Greed, fear and stock market dynamics. Physica A: Stat Mech Appl 343:635–642

Books and Reviews

Anderson PW, Arrow KJ, Pines D (1989) The Economy as an Evolving Complex System. Addison-Wesley, Reading

Arthur WB, Durlauf SN, Lane DA (1999) The Economy as an Evolving Complex System: vol 2. Addison-Wesley, Reading

Blume LE, Durlauf SN (2005) The Economy as an Evolving Complex System: vol 3. Oxford University Press, Oxford

Judd KL, Tesfatsion L (2006) Handbook of Computational Economics, vol 2: Agent-based Computational Economics. North-Holland, Amsterdam

Financial Economics, Fat-Tailed Distributions

Markus Haas[1], Christian Pigorsch[2]
[1] Department of Statistics, University of Munich, Munich, Germany
[2] Department of Economics, University of Bonn, Bonn, Germany

Article Outline

Glossary

Leptokurtosis A distribution is leptokurtic if it is more peaked in the center and thicker tailed than the normal distribution with the same mean and variance. Occasionally, leptokurtosis is also identified with a moment-based kurtosis measure larger than three, see Sect. "Introduction".

Return Let S_t be the price of a financial asset at time t. Then the *continuous* return, r_t, is $r_t = \log(S_t/S_{t-1})$. The *discrete* return, R_t, is $R_t = S_t/S_{t-1} - 1$. Both are rather similar if $-0.15 < R_t < 0.15$, because $r_t = \log(1 + R_t)$. See Sect. "Introduction".

Tail The (upper) tail, denoted by $\bar{F}(x) = P(X > x)$, characterizes the probability that a random variable X exceeds a certain "large" threshold x. For analytical purposes, "large" is often translated with "as $x \to \infty$". For financial returns, a daily change of 5% is already infinitely large. A Gaussian model essentially excludes such an event.

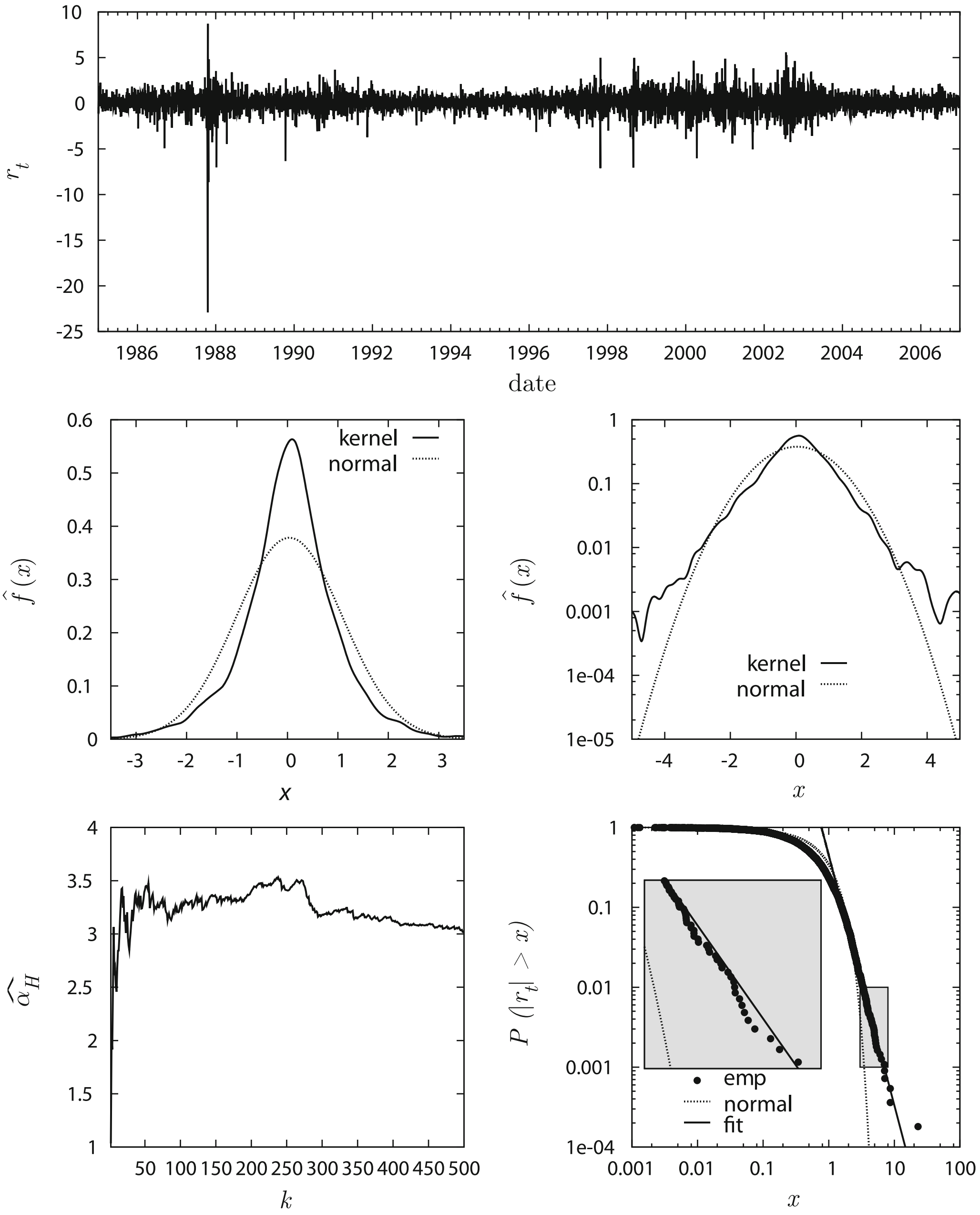
Tail index The tail index, or *tail exponent*, α, characterizes the rate of tail decay if the tail goes to zero, in essence, like a power function, i. e., $\bar{F}(x) = x^{-\alpha}L(x)$, where L is slowly varying. Moments of order lower (higher) than α are (in)finite.

Definition of the Subject

Have a look at Fig. 1. The top plot shows the daily percentage changes, or *returns*, of the S&P500 index ranging from January 2, 1985 to December 29, 2006, a total of 5,550 daily observations. We will use this data set throughout the article to illustrate some of the concepts and models to be discussed. Two observations are immediate. The first is that both small and large changes come clustered, i. e., there are periods of low and high volatility. The second is that, from time to time, we observe rather large changes which may be hard to reconcile with *the* standard distributional assumption in statistics and econometrics, that is, normality. The most outstanding return certainly occurred on October 19, 1987, the "Black Monday", where the index lost more than 20% of its value, but the phenomenon is chronic. For example, if we fit a normal distribution to the data, the resulting model predicts that we observe an absolute daily change larger than 5% once in approximately 1,860 years, whereas we actually encountered that 13 times during our 22-year sample period. This suggests that, compared to the normal distribution, the distribution of the returns is *fat-tailed*, i. e., the probability of large losses and gains is much higher than would be implied by a time-invariant unconditional Gaussian distribution. The latter is obviously not suitable for describing the booms, busts, bubbles, and bursts of activity which characterize financial markets, and which are apparent in Fig. 1.

The two aforementioned phenomena, i. e., volatility clustering and fat tails, have been detected in almost every financial return series that was subject to statistical analysis since the publication of Mandelbrot's [155] seminal study of cotton price changes, and they are of paramount importance for any individual or institution engaging in the financial markets, as well as for financial economists trying to understand their mode of operation. For example, investors holding significant portions of their wealth in risky assets need a realistic assessment of the likelihood of severe losses. Similarly, economists trying to learn about the relation between risk and return, the pricing of financial derivatives, such as options, and the inherent dynamics of financial markets, can only benefit from building their models on adequate assumptions about the stochastic properties of the variables under study, and they have to reconcile the predictions of their models with the actual facts.

This article reviews some of the most important concepts and distributional models that are used in empirical finance to capture the (almost) ubiquitous stochastic properties of returns as indicated above. Section "Introduction" defines in a somewhat more precise manner than above the central variable of interest, the return of a financial asset, and gives a brief account of the early history of the problem. Section "Defining Fat-Tailedness" discusses various operationalizations of the term "fat-tailedness",



Financial Economics, Fat-Tailed Distributions, Figure 1
The *top plot* shows the S&P500 *percentage* returns, r_t, from January 1985 to December 2006, i. e., $r_t = 100 \times \log(S_t/S_{t-1})$, where S_t is the index level at time t. The left plot of the *middle panel* shows a nonparametric density estimate (*solid*), along with the fitted normal density (*dotted*); the right graph is similar but shows the respective log-densities in order to better visualize the tail regions. The *bottom left plot* represents a *Hill plot* for the S&P500 returns, i. e., it displays $\hat{\alpha}_{k,n}$ defined in (11) for $k \leq 500$. The *bottom right plot* shows the complementary cdf, $\bar{F}(x)$, on a log-log scale, see Sect. "Empirical Evidence About the Tails" for discussion

and Sect. "Empirical Evidence About the Tails" summarizes what is or is at least widely believed to be known about the tail characteristics of typical return distributions. Popular parametric distributional models are discussed in Sect. "Some Specific Distributions". The alpha stable model as the archetype of a fat-tailed distribution in finance is considered in detail, as is the generalized hyperbolic distribution, which provides a convenient framework for discussing, as special or limiting cases, many of the important distributions employed in the literature. An empirical comparison using the S&P500 returns is also included. In Sect. "Volatility Clustering and Fat Tails", the relation between the two "stylized facts" mentioned above, i. e., clusters of volatility and fatness of the tails, is highlighted, where we concentrate on the GARCH approach, which has gained outstanding popularity among financial econometricians. This model has the intriguing property of producing fat-tailed marginal distributions even with light-tailed innovation processes, thus emphasizing the role of the market dynamics. In Sect. "Application to Value-at-Risk", we compare both the unconditional parametric distributional models introduced in Sect. "Some Specific Distributions" as well as the GARCH model of Sect. "Volatility Clustering and Fat Tails" on an economic basis by evaluating their ability to accurately measure the Value-at-Risk, which is an important tool in risk management. Finally, Sect. "Future Directions" identifies some open issues.

Introduction

To fix notation, let S_t be the price of an asset at time t, e. g., a stock, a market index, or an exchange rate. The *continuously compounded* or *log* return from time t to time $t + \Delta t$, $r_{t,t+\Delta t}$, is then defined as

$$r_{t,t+\Delta t} = \log S_{t+\Delta t} - \log S_t \, . \tag{1}$$

Often the quantity defined in (1) is also multiplied by 100, so that it can be interpreted in terms of *percentage returns*, see Fig. 1. Moreover, in applications, Δt is usually set equal to one and represents the horizon over which the returns are calculated, e. g., a day, week, or month. In this case, we drop the first subscript and define $r_t := \log S_t - \log S_{t-1}$. The log returns (1) can be additively aggregated over time, i. e.,

$$r_{t,t+\tau} = \sum_{i=1}^{\tau} r_{t+i} \, . \tag{2}$$

Empirical work on the distribution of financial returns is usually based on log returns. In some applications a useful fact is that, over short intervals of time, when returns tend to be small, (1) can also serve as a reasonable approximation to the *discrete* return, $R_{t,t+\Delta t} := S_{t+\Delta t}/S_t - 1 = \exp(r_{t,t+\Delta t}) - 1$. For further discussion of the relationship between continuous and discrete returns and their respective advantages and disadvantages, see, e. g., [46,76].

The seminal work of Mandelbrot [155], to be discussed in Subsect. "Alpha Stable and Related Distributions", is often viewed as the beginning of modern empirical finance. As reported in [74], "[p]rior to the work of Mandelbrot the usual assumption ... was that the distribution of price changes in a speculative series is approximately Gaussian or normal". The rationale behind this prevalent view, which was explicitly put forward as early as 1900 by Bachelier [14], was clearly set out in [178]: If the log-price changes (1) from transaction to transaction are independently and identically distributed with finite variance, and if the number of transactions is fairly uniformly distributed in time, then (2) along with the central limit theorem (CLT) implies that the return distribution over longer intervals, such as a day, a week, or a month, approaches a Gaussian shape.

However, it is now generally acknowledged that the distribution of financial returns over horizons shorter than a month is not well described by a normal distribution. In particular, the empirical return distributions, while unimodal and approximately symmetric, are typically found to exhibit considerable *leptokurtosis*, i. e., they are more peaked in the center and have fatter tails than the Gaussian with the same variance. Although this has been occasionally observed in the pre-Mandelbrot literature (e. g., [6]), the first systematic account of this phenomenon appeared in [155] and the follow-up papers by Fama [74,75] and Mandelbrot [156], and it was consistently confirmed since then. The typical shape of the return distribution, as compared to a fitted Gaussian, is illustrated in the middle panel of Fig. 1 for the S&P500 index returns, where a nonparametric kernel density estimator (e. g., [198]) is superimposed on the fitted Gaussian curve (dashed line). Interestingly, this pattern has been detected not only for modern financial markets but also for those of the eighteenth century [103].

The (location and scale-free) standardized fourth moment, or coefficient of *kurtosis*,

$$\mathbb{K}[X] = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4} \, , \tag{3}$$

where μ and σ are the mean and the standard deviation of the random variable (rv) X, respectively, is sometimes used to assess the degree of leptokurtosis of a given distribution. For the normal distribution, $\mathbb{K} = 3$, and $\mathbb{K} > 3$, referred to as *excess kurtosis*, is taken as an indicator of a leptokurtic shape (e. g., [164], p. 429). For example, the sample analogue of (3) for the S&P500 returns shown in Fig. 1 is 47.9, indicating very strong excess kurtosis. A formal test could be conducted using the fact that, under normality, the sample kurtosis is asymptotically normal with mean 3 and standard deviation $\sqrt{24/T}$ (T being the sample size), but the result can be anticipated.

As is well-known, however, such moment-based summary measures have to be interpreted with care, because a particular moment need not be very informative about a density's shape. We know from Finucan [82] that if two symmetric densities, f and g, have common mean and variance and finite fourth moment, and if g is more peaked and has thicker tails than f, *then* the fourth moment (and hence $\mathbb{K}$) is greater for g than for f, provided the densities cross exactly twice on both sides of the mean. However, the converse of this statement is, of course, not true, and a couple of (mostly somewhat artificial) counterexamples can be found in [16,68,121]. [158] provides some intuition by relating density crossings to moment crossings. For example, (only) if the densities cross more than four times, it may happen that the fourth moment is greater for f, but the sixth and all higher moments are greater for g, reflecting the thicker tails of the latter. Nevertheless, Finucan's result, along with his (in some respects justified) hope that we can view "this pattern as the common explanation of a high observed kurtosis", may serve to argue for a certain degree of usefulness of the kurtosis measure (3), provided the fourth moment is assumed to be finite. However, a nonparametric density estimate will in any case be more informative. Note that the density crossing condition in Finucan's theorem is satisfied for the S&P500 returns in Fig. 1.

Defining Fat-Tailedness

The notion of leptokurtosis as discussed so far is rather vague, and both financial market researchers as well as practitioners, such as risk managers, are interested in a more precise description of the tail behavior of financial variables, i. e., the laws governing the probability of large gains and losses. To this end, we define the *upper tail* of the distribution of a rv X as

$$\bar{F}(x) = P(X > x) = 1 - F(x) , \tag{4}$$

where F is the cumulative distribution function (cdf) of X. Consideration of the upper tail is the standard convention in the literature, but it is clear that everything could be phrased just as well in terms of the lower tail.

We are interested in the behavior of (4) as x becomes large. For our benchmark, i. e., the normal distribution with (standardized) density (pdf) $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, we have (cf. p. 131 in [79])

$$\bar{F}(x) \cong \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right) = \frac{\phi(x)}{x} \quad \text{as } x \to \infty , \tag{5}$$

where the notation $f(x) \cong g(x)$ as $x \to \infty$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Thus, the tails of the normal tend to zero faster than exponentially, establishing its very light tails.

To appreciate the difference between the general concept of leptokurtosis and the approach that focuses on the tails, consider the class of finite normal mixtures as discussed in Subsect. "Finite Mixtures of Normal Distributions". These are leptokurtic in the sense of peakedness and tailedness (compared to the normal), but are light-tailed according to the tail-based perspective.

While it is universally accepted in the literature that the Gaussian is too light-tailed to be an appropriate model for the distribution of financial returns, there is no complete agreement with respect to the actual shape of the tails. This is not surprising because we cannot reasonably expect to find a model that accurately fits all markets at any time and place. However, the current mainstream opinion is that the probability for the occurrence of large (positive and negative) returns can often appropriately be described by Pareto-type tails. Such tail behavior is also frequently adopted as the definition of fat-tailedness per se, but the terminology in the literature is by no means unique.

A distribution has Pareto-type tails if they decay essentially like a power function as x becomes large, i. e., $\bar{F}$ is regularly varying (at infinity) with index $-\alpha$ (written $\bar{F} \in \mathrm{RV}_{-\alpha}$), meaning that

$$\bar{F}(x) = x^{-\alpha} L(x) , \quad \alpha > 0 , \tag{6}$$

where $L > 0$ is a slowly varying function, which can be interpreted as "slower than any power function" (see [34, 188,195] for a technical treatment of regular variation). The defining property of a slowly varying function is $\lim_{x\to\infty} L(tx)/L(x) = 1$ for any $t > 0$, and the aforementioned interpretation follows from the fact that, for any $\gamma > 0$, we have (cf. [195], p. 18)

$$\lim_{x\to\infty} x^{\gamma} L(x) = \infty , \quad \lim_{x\to\infty} x^{-\gamma} L(x) = 0 . \tag{7}$$

Thus, for large x, the parameter α in (6), called the *tail index* or *tail exponent*, controls the rate of tail decay and provides a measure for the fatness of the tails.

Typical examples of slowly varying functions include $L(x) = c$, a constant, $L(x) = c + o(1)$, or $L(x) = (\log x)^k$, $x > 1$, $k \in \mathbb{R}$. The first case corresponds to strict Pareto tails, while in the second the tails are asymptotically Paretian in the sense that $\bar{F}(x) \cong cx^{-\alpha}$, which includes as important examples in finance the (non-normal) stable Paretian (see (13) in Subsect. "Alpha Stable and Related Distributions") and the Student's t distribution considered in Sect. "The Student t Distribution", where the tail index coincides with the characteristic exponent and the number of degrees of freedom, respectively. As an example for both, the Cauchy distribution with density $f(x) = [\pi(1+x^2)]^{-1}$ has cdf $F(x) = 0.5 + \pi^{-1} \arctan(x)$.

As $\arctan(x) = \sum_0^\infty (-1)^i x^{2i+1}/(2i+1)$ for $|x| < 1$, and $\arctan(x) = \pi/2 - \arctan(1/x)$ for $x > 0$, we have $\bar{F}(x) \cong (\pi x)^{-1}$.

For the distributions mentioned in the previous paragraph, it is known that their moments exist only up to (and excluding) their tail indices, α. This is generally true for rvs with regularly varying tails and follows from (7) along with the well-known connection between moments and tail probabilities, i. e., for a non-negative rv X, and $r > 0$, $\mathbb{E}[X^r] = r\int_0^\infty x^{r-1}\bar{F}(x)\mathrm{d}x$ (cf. [95], p. 75). The only possible minor variation is that, depending on L, $\mathbb{E}[X^\alpha]$ may be finite. For example, a rv X with tail $\bar{F}(x) = cx^{-1}(\log x)^{-2}$ has finite mean. The property that moments greater than α do not exist provides further intuition for α as a measure of tail-fatness.

As indicated above, there is no universal agreement in the literature with respect to the definition of fat-tailedness. For example, some authors (e. g., [72,196]) emphasize the class of *subexponential* distributions, which are (although not exclusively) characterized by the property that their tails tend to zero slower than any exponential, i. e., for any $\gamma > 0$, $\lim_{x\to\infty} \mathrm{e}^{\gamma x}\bar{F}(x) = \infty$, implying that the moment generating function does not exist. Clearly a regularly varying distribution is also subexponential, but further members of this class are, for instance, the lognormal as well as the *stretched exponential*, or heavy-tailed Weibull, which has a tail of the form

$$\bar{F}(x) = \exp\left(-x^b\right), \quad 0 < b < 1. \tag{8}$$

The stretched exponential has recently been considered by [134,152,153] as an alternative to the Pareto-type distribution (6) for modeling the tails of asset returns. Note that, as opposed to (6), both the lognormal as well as the stretched exponential possess power moments of all orders, although no exponential moment.

In addition, [22] coined the term *semi-heavy tails* for the generalized hyperbolic (GH) distribution, but the label may be employed more generally to refer to distributions with slower tails than the normal but existing moment generating function. The GH, which is now very popular in finance and nests many interesting special cases, will be examined in detail in Subsect. "The Generalized Hyperbolic Distribution".

As will be discussed in Sect. "Empirical Evidence About the Tails", results of extreme value theory (EVT) are often employed to identify the tail shape of return distributions. This has the advantage that it allows one to concentrate fully on the tail behavior, without the need to model the central part of the distribution. To sketch the idea behind this approach, suppose we attempt to classify distributions according to the limiting behavior of their normalized maxima. To this end, let $\{X_i, i \geq 1\}$ be an iid sequence of rvs with common cdf F, $M_n = \max\{X_1, \ldots, X_n\}$, and assume there exist sequences $a_n > 0$, $b_n \in \mathbb{R}$, $n \geq 1$, such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \xrightarrow{n\to\infty} G(x), \tag{9}$$

where G is assumed nondegenerate. To see that normalization is necessary, note that $\lim_{n\to\infty} P(M_n \leq x) = \lim_{n\to\infty} F^n(x) = 0$ for all $x < x_M := \sup\{x : F(x) < 1\} \leq \infty$, so that the limiting distribution is degenerate and of little help. If the above assumptions are satisfied, then, according to the classical Fisher–Tippett theorem of extreme value theory (cf. [188]), the limiting distribution G in (9) must be of the following form:

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-1/\xi}\right), \quad 1 + \xi x > 0, \tag{10}$$

which is known as the *generalized extreme value distribution* (GEV), or *von Mises representation of the extreme value distributions* (EV). The latter term can be explained by the fact that (10) actually nests three different types of EV distributions, namely

(i) the Fréchet distribution, denoted by G_ξ^+, where $\xi > 0$ and $x > -1/\xi$,
(i) the so-called Weibull distribution of EVT, denoted by G_ξ^-, where $\xi < 0$ and $x < -1/\xi$, and
(iii) the Gumbel distribution, denoted by G_0, which corresponds to the limiting case as $\xi \to 0$, i. e., $G_0(x) = \exp(-\exp(-x))$, where $x \in \mathbb{R}$.

A cdf F belongs to the *maximum domain of attraction* (MDA) of one of the extreme value distributions nested in (10), written $F \in \mathrm{MDA}(G_\xi)$, if (9) holds, i. e., classifying distributions according to the limiting behavior of their extrema amounts to figuring out the MDAs of the extreme value distributions. It turns out that it is the tail behavior of a distribution F that accounts for the MDA it belongs to. In particular, $F \in \mathrm{MDA}(G_\xi^+)$ if and only if its tail $\bar{F} \in \mathrm{RV}_{-\alpha}$, where $\alpha = 1/\xi$. As an example, for a strict Pareto distribution, i. e., $F(x) = 1 - (u/x)^\alpha$, $x \geq u > 0$, with $a_n = n^{1/\alpha}u/\alpha$ and $b_n = n^{1/\alpha}u$, we have $\lim_{n\to\infty} F^n(a_n x + b_n) = \lim_{n\to\infty}(1 - n^{-1}(1 + x/\alpha)^{-\alpha})^n = G_{1/\alpha}^+(x)$. Distributions in $\mathrm{MDA}(G_\xi^-)$ have a finite right endpoint, while, roughly, most of the remaining distributions, such as the normal, the lognormal and (stretched) exponentials, belong to $\mathrm{MDA}(G_0)$. The latter also accommodates a few distributions with finite right endpoint. See [188] for precise conditions. The important case of non-iid rvs is discussed in [136]. A central result is that, rather generally, vis-à-vis an iid sequence with the same marginal cdf, the maxima of stationary sequences converge to the same type of limiting distribution. See [63,167] for an application of this theory to ARCH(1) and GARCH(1,1) processes (see Sect. "Volatility Clustering and Fat Tails"), respectively.

One approach to exploit the above results, referred to as the *method of block maxima*, is to divide a given sample of return data into subsamples of equal length, pick the maximum of each subsample, assume that these have been generated by (10) (enriched with location and scale parameters to account for the unknown a_n and b_n), and find the maximum-likelihood estimate for ξ, location, and scale. Standard tests can then be conducted to assess, e. g., whether $\xi > 0$, i. e., the return distribution has Pareto-type tails. An alternative but related approach, which is based on further theoretical developments and often makes more efficient use of the data, is the *peaks over thresholds* (POT) method. See [72] for a critical discussion of these and alternative techniques.

We finally note that $1 - G^{+}_{1/\alpha}(\alpha(x-1)) \cong x^{-\alpha}$, while $1 - G_0(x) \cong \exp(-x)$, i. e., for the extremes, we have asymptotically a Pareto and an exponential tail, respectively. This may provide, on a meta-level, a certain rationale for reserving the notion of genuine fat-tailedness for the distributions with regularly varying tails.

Empirical Evidence About the Tails

The first application of power tails in finance appeared in Mandelbrot's [155] study of the log-price changes of cotton. Mandelbrot proposed to model returns with non-normal *alpha stable*, or *stable Paretian*, distributions, the properties of which will be discussed in some detail in Subsect. "Alpha Stable and Related Distributions". For the present discussion, it suffices to note that for this model the tail index α in (6), also referred to as *characteristic exponent* in the context of stable distributions, is restricted to the range $0 < \alpha < 2$, and that much of its theoretical appeal derives from the fact that, due to the generalized CLT, "Mandelbrot's hypothesis can actually be viewed as a generalization of the central-limit theorem arguments of Bachelier and Osborne to the case where the underlying distributions of price changes from transaction to transaction … have infinite variances" [75]. For the cotton price changes, Mandelbrot came up with a tail index of about 1.7, and his work was subsequently complemented by Fama [75] with an analysis of daily returns of the stocks belonging to the Dow Jones Industrial Average. [75] came to the conclusion that Mandelbrot's theory was supported by these data, with an average estimated α close to 1.9.

The findings of Mandelbrot and Fama initiated an extensive discussion about the appropriate distributional model for stock returns, partly because the stable model's implication that the tails are so fat that even the variance is infinite appeared to be too radical to many economists used to working with models built on the assumption of finite second moments. The evidence concerning the stable hypothesis gathered in the course of the debate until the end of the 1980s was not ultimately conclusive, but there were many papers reporting mainly negative results [4,28,36,40,54,67,98,99,109,135,176,180,184].

From the beginning of the 1990s, a number of researchers have attempted to estimate the tail behavior of asset returns directly, i. e., without making specific assumptions about the entire distributional shape. [86,115,142,143] use the method of block maxima (see Sect. "Defining Fat-Tailedness") to identify the maximum domain of attraction of the distribution of stock returns. They conclude that the Fréchet distribution with a tail index $\alpha > 2$ is most likely, implying Pareto-type tails which are thinner than those of the *stable* Paretian.

A second strand of literature assumes a priori the presence of a Pareto-type tail and focuses on the estimation of the tail index α. If, as is often the case, a power tail is deemed adequate, an explicit estimate of α is of great interest both from a practical and an academic viewpoint. For example, investors want to assess the likelihood of large losses of financial assets. This is often done using methods of extreme value theory, which require an accurate estimate of the tail exponent. Such estimates are also important because the properties of statistical tests and other quantities of interest, such as empirical autocorrelation functions, frequently depend on certain moment conditions (e. g., [144,167]). Clearly the desire to figure out the actual tail shape has also an intrinsic component, as is reflected in the long-standing debate on the stable Paretian hypothesis. People simply wanted to know whether this distribution, with its appealing theoretical properties, is consistent with actual data. Moreover, empirical findings may guide economic theorizing, as they can help both in assessing the validity of certain existing models as well as in suggesting new explanations. Two examples will briefly be discussed at the end of the present section.

Within this second strand of literature, the Hill estimator [106] has become the most popular tool. It is given by

$$\hat{\alpha}_{k,n} = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log X_{j,n} - \log X_{k,n} \right)^{-1}, \tag{11}$$

where $X_{i,n}$ denotes the ith upper order statistic of a sample of length n, i. e., $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{n,n}$. See [64, 72] for various approaches to deriving (11). If the tail is not regularly varying, the Hill estimator does not estimate anything.

A crucial choice to be made when using (11) is the selection of the threshold value k, i. e., the number of order statistics to be included in the estimation. Ideally, only observations from the tail region should be used, but choosing k to small will reduce the precision of the estimator. There exist various methods for picking k optimally in a mean-squared error sense [61,62], but much can be learned by looking at the *Hill plot*, which is obtained by plotting $\hat{\alpha}_{k,n}$ against k. If we can find a range of k-values where the estimate is approximately constant, this can be taken as a hint for where the "true" tail index may be located. As illustrated in [189], however, the Hill plot may not always be so well-behaved, and in this case the semi-automatic methods mentioned above will presumably also be of little help.

The theoretical properties of (11), along with technical conditions, are summarized in [72,189]. Briefly, for iid data generated from a distribution with tail $\bar{F} \in \mathrm{RV}_{-\alpha}$, the Hill estimator has been shown to be consistent [159] and asymptotically normal with standard deviation $\alpha/\sqrt{k}$ [100]. Financial data, however, are usually not iid but exhibit considerable dependencies in higher-order moments (see Sect. "Volatility Clustering and Fat Tails"). In this situation, i. e., with ARCH-type dynamics, (11) will still be consistent [190], but little is known about its asymptotic variance. However, simulations conducted in [123] with an IGARCH model, as defined in Sect. "Volatility Clustering and Fat Tails", indicate that, under such dependencies, the actual standard errors may be seven to eight times larger than those implied by the asymptotic theory for iid variables.

The Hill estimator was introduced into the econometrics literature in the series of articles [107,113,125,126]. [125,126], using weekly observations, compare the tails of exchange rate returns in floating and fixed exchange rate systems, such as the Bretton Woods period and the EMS. They find that for the fixed systems, most tail index estimates are below 2, i. e., consistent with the alpha stable hypothesis, while the estimates are significantly larger than 2 (ranging approximately from 2.5 to 4) for the float. [126] interpret these results in the sense that "a float lets exchange rates adjust more smoothly than any other regime that involves some amount of fixity". Subsequent studies of floating exchange rates using data ranging from weekly [107,110,111] over daily [58,89,144] to very high-frequency [59,61,170] have confirmed the finding of these early papers that the tails are not fat enough to be compatible with the stable Paretian hypothesis, with estimated tail indices usually somewhere in the region 2.5–5. [58] is the first to investigate the tail behavior of the euro against the US dollar, and finds that it is similar both to the German mark in the pre-euro era as well as to the yen and the British pound, with estimated exponents hovering around 3–3.5.

Concerning estimation with data at different time scales, a comparison of the results reported in the literature reveals that the impact on the estimated tail indices appears to be moderate. [59] observe an increase in the estimates when moving from 30-minute to daily returns, but they argue that these changes, due to the greater bias at the lower frequencies, are small enough to be consistent with α being invariant under time aggregation.

Note that if returns were independently distributed, their tail behavior would in fact be unaffected by time aggregation. This is a consequence of (2) along with Feller's (p. 278 in [80]) theorem on the convolution of regularly varying distributions, stating that any finite convolution of a regularly varying cdf $F(x)$ has a regularly varying tail with the same index. Thus, in principle, the tail survives forever, but, as long as the variance is finite, the CLT ensures that in the course of aggregation an increasing probability weight is allocated to the center of the distribution, which becomes closer to a Gaussian shape. The probability of observing a tail event will thus decrease. However, for fat-tailed distributions, the convergence to normality can be rather slow, as reflected in the observation that pronounced non-normalities in financial returns are often observed even at a weekly and (occasionally) monthly frequency. See [41] for an informative discussion of these issues. The fact that returns are, in general, not iid makes the interpretation of the approximate stability of the tail index estimates observed across papers employing different frequencies not so clear-cut, but Feller's theorem may nevertheless provide some insight.

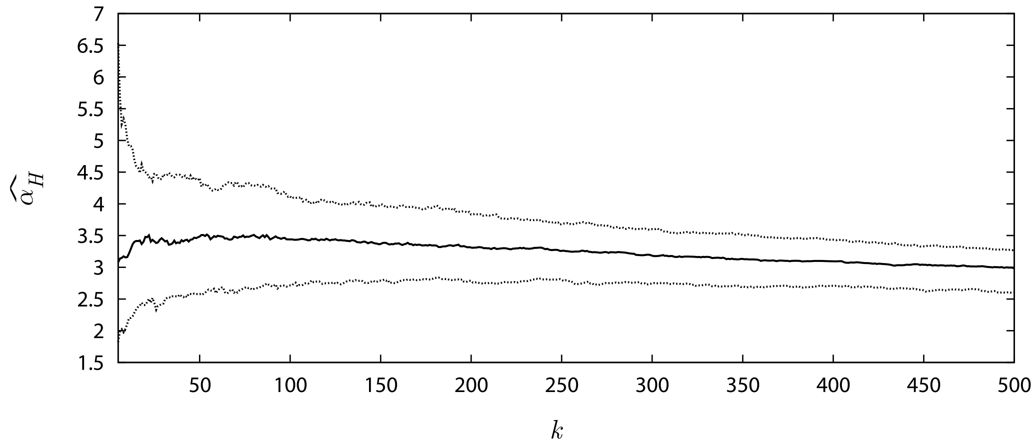
There is also an extensive literature reporting tail index estimates of stock returns, mostly based on daily [2,89,92, 112,113,144,145,146,177] and higher frequencies [2,91,92, 147,181]. The results are comparable to those for floating exchange rates in that the tenor of this literature, which as a whole covers all major stock markets, is that most stock return series are characterized by a tail index somewhere in the region 2.5–5, and most often below 4. That is, the tails are thinner than expected under the stable Paretian hypothesis, but the finiteness of the third and in particular the fourth moments (and hence kurtosis) may already be questionable. Again, the results appear to be relatively insensitive with respect to the frequency of the data, indicating a rather slow convergence to the normal distribution. Moreover, most authors do not find significant differences between the left and the right tail, although, for stock returns, the point estimates tend to be somewhat lower for the left tail (e. g., [115,145]).

Applications to the bond market appear to be rare, but see [201], who report tail index estimates between 2.5 and 4.5 for 5-minute and 1-hour Bund future returns and somewhat higher values for daily data. [160] compare the tail behaviors of spot and future prices of various commodities (including cotton) and find that, while future prices resemble stock prices with tail indices in the region 2.5–4, spot prices are somewhat fatter tailed with α hovering around 2.5 and, occasionally, smaller than 2.

Summarizing, it is now a widely held view that the distribution of asset returns can typically be described as fat-tailed in the power law sense but with finite variance. Thus, currently there seems to exist a far reaching consensus that the stable Paretian model is not adequate for financial data, but see [162,202] for a different viewpoint. A consequence of the prevalent view is that asset return distributions belong to the Gaussian domain of attraction, but that the convergence appears to be very slow.

To illustrate typical findings as reported above, let us consider the S&P500 returns described in Sect. “Definition of the Subject”. A first informal check of the appropriateness of a power law can be obtained by means of a log-log plot of the empirical tail, i. e., if $1 - F(x) = \bar{F}(x) \approx cx^{-\alpha}$ for large x, then a plot of the log of the empirical complementary cdf, $\bar{F}(x)$, against $\log x$ should display a linear behavior in its outer part. For the data at hand, such a plot is shown in the bottom right panel of Fig. 1. Assuming homogeneity across the tails, we pool negative and positive returns by first removing the sample mean and then taking absolute values. We have also multiplied (1) by 100, so that the returns are interpretable in terms of percentage changes. The plot suggests that a power law regime may be starting from approximately the 90% quantile. Included in Fig. 1 is also a regression line (“fit”) fitted to the log-tail using the 500 upper (absolute) return observations. This yields, as a rough estimate for the tail index, a slope of $\hat{\alpha} = 3.13$, with a coefficient of determination $R^2 = 0.99$. A Hill plot for $k \leq 500$ in (11) is shown in the bottom left panel of Fig. 1. The estimates are rather stable over the entire region and suggest an α somewhere in the interval $(3, 3.5)$, which is reconcilable with the results in the literature summarized above. A somewhat broader picture can be obtained by considering individual stocks. Here we consider the 176 stocks that were listed in the S&P500 from January 1985 to December 2006. Figure 2 presents, for each $k \leq 500$, the 5%, 50%, and 95% quantiles of the distribution of (11) over the different stocks. The median is close to 3 throughout, and it appears that an estimate in $(2.5, 4.5)$ would be reasonable for most stocks.

At this point, it may be useful to note that the issue is not whether a power law is true in the strict sense but only if it provides a reasonable approximation in the relevant tail region. For example, it might be argued that asset returns actually have finite support, implying finiteness of all moments and hence inappropriateness of a Pareto-type tail. However, as concisely pointed out in [144], “saying that the support of an empirical distribution is bounded says very little about the nature of outlier activity that may occur in the data”.

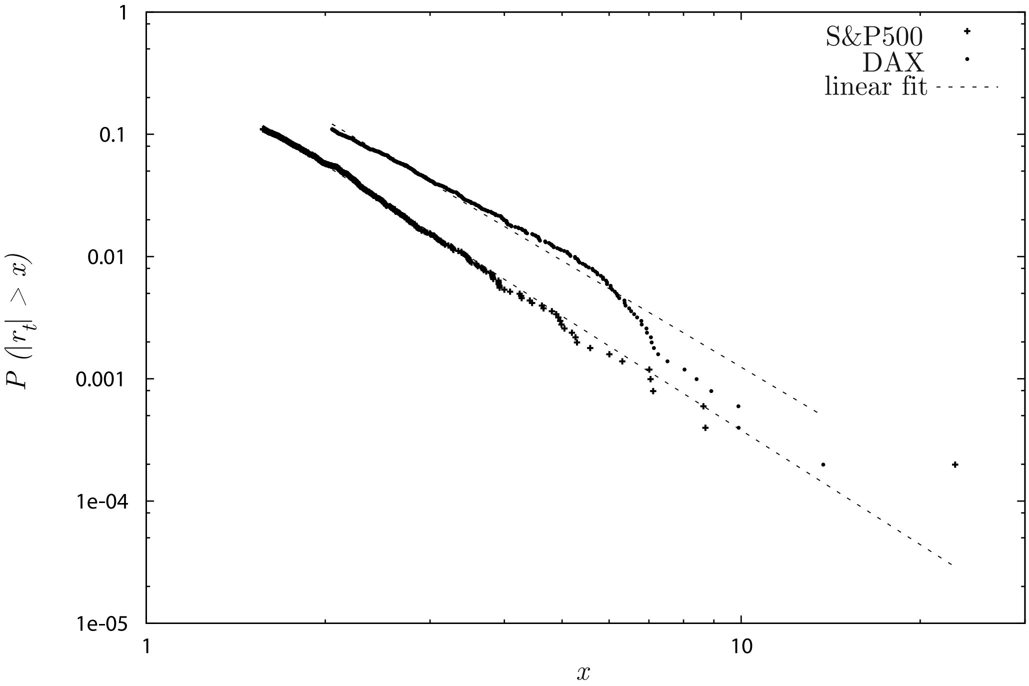


Financial Economics, Fat-Tailed Distributions, Figure 2
Shown are, for $k \leq 500$, the 95%, 50%, and 5% quantiles of the distribution of the Hill estimator $\hat{\alpha}_{k,n}$, as defined in (11), over 176 stocks included in the S&P500 stock index

We clearly cannot expect to identify the "true" distribution of financial variables. For example, [153] have demonstrated that by standard techniques of EVT it is virtually impossible, even in rather large samples, to discriminate between a power law and a stretched exponential (8) with a small value of b, thus questioning, for example, the conclusiveness of studies relying on the block maxima method, as referred to above. A similar point was made in [137], who showed by simulation that a three-factor stochastic volatility model, with a marginal distribution known to have all its moments finite, can generate apparent power laws in practically relevant sample sizes. As put forward in [152], "for most practical applications, the relevant question is not to determine what is the true asymptotic tail, but what is the best effective description of the tails in the domain of useful applications".

As is evident in Fig. 1, a power law may (and often does) provide a useful approximation to the tail behavior of actual data, but there is no reason to expect that it will appear in every market, and a broad range of heavy and semi-heavy tailed distributions (such as the GH in Subsect. "The Generalized Hyperbolic Distribution") may provide an adequate fit. For instance, [93] investigate the tail behavior of high-frequency returns of one of the most frequently traded stocks on the Paris Stock Exchange (Alcatel) and conclude that the tails decay at an exponential rate, and [119,197] obtain similar results for daily returns of the Nikkei 225 index and various individual US stocks, respectively. As a further illustration, without rigorous statistical testing, Fig. 3 shows the log-log tail plot for daily returns of the German stock market index DAX from July 3, 1987 to July 4, 2007 for the largest 500 out of 5,218 (absolute) return observations, along with a regression-based linear fit. For purposes of comparison, the corresponding figure for the S&P500 has also been reproduced from Fig. 1. While the slopes of the fitted power laws exhibit an astonishing similarity (in fact, the estimated tail index of the DAX is 2.93), it is clear from Fig. 3 that an asymptotic power law, although not necessarily inconsistent with the data, is much less self-evident for the DAX than for the S&P500, due to the apparent curvature particularly in the more extreme tail.

It is finally worthwhile to mention that financial theory in general, although some of its models are built on the *assumption* of a specific distribution, has little to say about the distribution of financial variables. For example, according to the efficient markets paradigm, asset prices change in response to the arrival of relevant new infor-



Financial Economics, Fat-Tailed Distributions, Figure 3
The figure shows, on a log-log scale, the complementary cdf, $\bar{F}(x)$, for the largest 500 absolute return observations both for the daily S&P500 returns from January 1985 to December 2006 and the daily DAX returns from July 1987 to July 2007

mation, and, consequently, the distributional properties of returns will essentially reflect those of the news process. As noted by [148], an exception to this rule is the model of rational bubbles of [35]. [148] show that this class of processes gives rise to an (approximate) power law for the return distribution. However, the structure of the model, i. e., the assumption of rational expectations, restricts the tail exponent to be below unity, which is incompatible with observed tail behaviors.

More recently, prompted by the observation that estimated tail indices are often located in a relatively narrow interval around 3, [83,84,85] have developed a model to explain a hypothesized "inverse cubic law for for the distribution of stock price variations" [91], valid for highly developed economies, i. e., a power law tail with index $\alpha = 3$. This model is based on Zipf's law for the size of large institutional investors and the hypothesis that large price movements are generated by the trades of large market participants via a square-root price impact of volume, V, i. e., $r \cong h\sqrt{V}$, where r is the log return and h is a constant. Putting these together with a model for profit maximizing large funds, which have to balance between trading on a perceived mispricing and the price impact of their actions, leads to a power law distribution of volume with tail index 1.5, which by the square-root price impact function and simple power law accounting then produces the "cubic law". See [78,182] for a discussion of this model and the validity of its assumptions. In a somewhat similar spirit, [161] find strong evidence for *exponentially* decaying tails of daily Indian stock returns and speculate about a general inverse relationship between the stage of development of an economy and the closeness to Gaussianity of its stock markets, but it is clear that this is really just speculation.

Some Specific Distributions

Alpha Stable and Related Distributions

As noted in Sect. "Empirical Evidence About the Tails", the history of heavy tailed distributions in finance has its origin in the *alpha stable* model proposed by Mandelbrot [154,155]. Being the first alternative to the Gaussian law, the alpha stable distribution has a long history in financial economics and econometrics, resulting in a large number of books and review articles.

Apart from its good empirical fit the stable distribution has also some attractive theoretical properties such as the stability property and domains of attraction. The stability property states that the index of stability (or shape parameter) remains the same under scaling and addition of different stable rv with the same shape parameter. The concept of domains of attraction is related to a generalized CLT. More specifically, dropping the assumption of a finite variance in the classical CLT, the domains of attraction states, loosely speaking, that the alpha stable distribution is the only possible limit distribution. For a more detailed discussion of this concept we refer to [169], who also provide an overview over alternative stable schemes. While the fat-tailedness of the alpha stable distributions makes it already an attractive candidate for modeling financial returns, the concept of the domains of attraction provides a further argument for its use in finance, as under the relaxation of the assumption of a finite variance of the continuously arriving return innovations the resulting return distribution at lower frequencies is generally an alpha stable distribution.

Although the alpha stable distribution is well established in financial economics and econometrics, there still exists some confusion about the naming convention and parameterization. Popular terms for the alpha stable distribution are the *stable Paretian*, *Lévy stable* or simply *stable* laws. The parameterization of the distribution in turn varies mostly with its application. For instance, to numerically integrate the characteristic function, it is preferable to have a continuous parameterization in all parameters.

The numerical integration of the alpha stable distributions is important, since with the exception of a few special cases, its pdf is unavailable in closed form. However, the characteristic function of the standard parameterization is given by

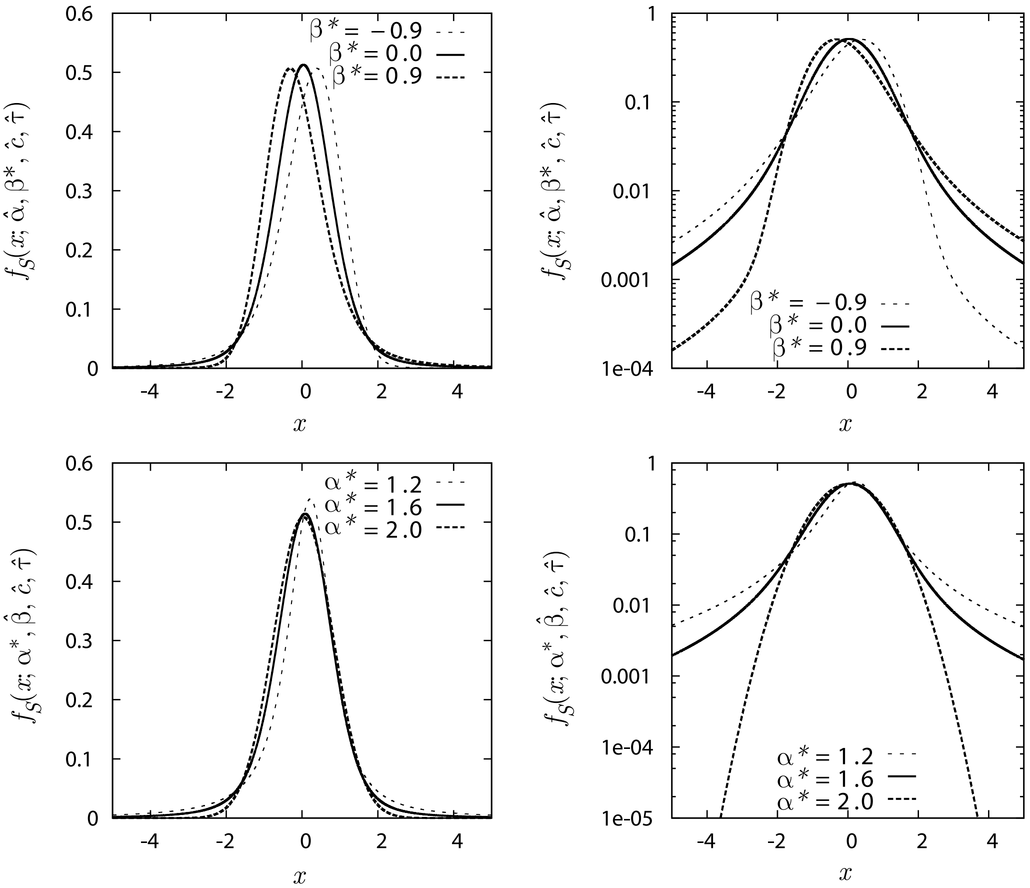
$$\mathbb{E}\left[\exp(itX)\right] = \begin{cases} \exp\Big(-c^{\alpha}\,|t|^{\alpha}\,\big(1 - i\beta\,\mathrm{sign}\,(t)\tan\tfrac{\pi\alpha}{2}\big) \\ \qquad\qquad + i\tau t\Big) \qquad \alpha \neq 1 \\ \exp\Big(-c\,|t|\,\big(1 + i\beta\tfrac{2}{\pi}\,\mathrm{sign}\,(t)\ln(|t|)\big) \\ \qquad\qquad + i\tau t\Big) \qquad \alpha = 1\,, \end{cases} \tag{12}$$

where i is the imaginary unit, sign $(\cdot)$ denotes the sign function, which is defined as

$$\mathrm{sign}\,(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0\,, \end{cases}$$

$0 < \alpha \leq 2$ denotes the *shape parameter*, *characteristic exponent* or *index of stability*, $-1 \leq \beta \leq 1$ is the skewness parameter, and $\tau \in \mathbb{R}$ and $c \geq 0$ are the location and scale parameters, respectively.

Figure 4 highlights the impact of the parameters α and β. β controls the skewness of the distribution. The shape parameter α controls the behavior of the tails of the distribution and therefore the degree of leptokurtosis. For $\alpha < 2$ moments only up to (and excluding) α exist, and for $\alpha > 1$ we have $\mathbb{E}[X] = \tau$. In general, for $\alpha \in (0, 1)$ and



Financial Economics, Fat-Tailed Distributions, Figure 4

Density function (pdf) of the alpha stable distribution for different parameter vectors. The *right panel* plots the log-densities to better visualize the tail behavior. The *upper (lower) section* present the pdf for different values of $\boldsymbol{\beta}$ ($\boldsymbol{\alpha}$)

$\beta = 1$ ($\beta = -1$) the support of the distribution is the set (τ, ∞) (or $(-\infty, \tau)$) rather than the whole real line. In the following we call this stable distribution with $\alpha \in (0, 1)$, $\tau = 0$ and $\beta = 1$ the positive alpha stable distribution.

Moreover, for $\alpha < 2$ the stable law has asymptotic power tails,

$$\bar{F}(x) = P(X > x) \cong c^{\alpha} d(1+\beta) x^{-\alpha}$$
$$f_S(x, \alpha, \beta, c, \tau) \cong \alpha c^{\alpha} d(1+\beta) x^{-\alpha+1}$$

with $d = \sin\left(\frac{\pi\alpha}{2}\right) \Gamma(\alpha)/\pi$.

For $\alpha = 2$ the stable law is equivalent to the normal law with variance $2c^2$, for $\alpha = 1$ and $\beta = 0$ the Cauchy distribution is obtained, and for $\alpha = 1/2$, $\beta = 1$ and $\tau = 0$ the stable law is equivalent to the Lévy distribution, with support over the positive real line.

An additional property of the stable laws is that they are closed under convolution for constant α, i. e., for two independent alpha stable rvs $X_1 \sim S(\alpha, \beta_1, c_1, \tau_1)$ and $X_2 \sim S(\alpha, \beta_2, c_2, \tau_2)$ with common shape parameter α we have

$$X_1 + X_2 \sim S\left(\alpha, \frac{\beta_1 c_1^{\alpha} + \beta_2 c_2^{\alpha}}{c_1^{\alpha} + c_2^{\alpha}}, \left(c_1^{\alpha} + c_2^{\alpha}\right)^{1/\alpha}, \tau_1 + \tau_2\right)$$

and

$$aX_1 + b \sim \begin{cases} S\left(\alpha, \operatorname{sign}(a)\beta, |a|c, a\tau + b\right) & \alpha \neq 1 \\ S\left(1, \operatorname{sign}(a)\beta, |a|c, a\tau + b - \frac{2}{\pi}\beta c a \log|a|\right) & \\ & \alpha = 1\,. \end{cases}$$

These results can be extended to n stable rvs. The closedness under convolution immediately implies the infinite divisibility of the stable law. As such every stable law corresponds to a Lévy process. A more detailed analysis of alpha stable processes in the context of Lévy processes is given in [192,193].

The computation and estimation of the alpha stable distribution is complicated by the aforementioned non-existence of a closed form pdf. As a consequence, a number of different approximations for evaluating the density have been proposed, see e. g. [65,175]. On the basis of these approximations, parameter estimation is facilitated using for example the maximum-likelihood estimator, see [66], or other estimation methods. As maximum-likelihood estimation relies on computationally demanding numerical integration methods, the early literature focused on alternative estimation methods. The most important methods include the quantile estimation suggested by [77,163], which is still heavily applied in order to obtain starting values for more sophisticated estimation procedures, as well as the characteristic function approach proposed by [127, 131,186]. However, based on its nice asymptotic properties and presently available computational power, the maximum-likelihood approach is preferable.

Many financial applications also involve the simulation of a return series. In derivative pricing, for example, the computation of an expectation is oftentimes impossible as the financial instrument is generally a highly nonlinear function of asset returns. A common way to alleviate this problem is to apply Monte Carlo integration, which in turn requires quasi rvs drawn from the respective return distribution, i. e. the alpha stable distribution. A useful simulation algorithm for alpha stable rvs is proposed by [49], which is a generalization of the algorithm of [120] to the non-symmetric case. A random variable X distributed according to the stable law, $S(\alpha, \beta, c, \tau)$, can be generated as follows:

1. Draw a rv U, uniformly distributed over the interval $(-\pi/2, \pi/2)$, and an (independent) exponential rv E with unit mean,
2. if $\alpha \neq 1$, compute

$$X = cS\frac{\sin(\alpha(U+B))}{\cos^{1/\alpha}(U)} \cdot \left(\frac{\cos(U-\alpha(U+B))}{E}\right)^{(1-\alpha)/\alpha} + \tau$$

where

$$B := \frac{\arctan\left(\beta \tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}$$

$$S := \left(1+\beta^2\tan^2\left(\frac{\pi\alpha}{2}\right)\right)^{1/(2\alpha)}$$

for $\alpha = 1$ compute

$$X = c\tfrac{2}{\pi}\left(\left(\tfrac{\pi}{2}+\beta U\right)\tan(U) - \beta\log\left(\frac{\frac{\pi}{2}E\cos(U)}{\frac{\pi}{2}+\beta U}\right)\right) + \frac{2}{\pi}\beta c\log(c) + \tau\,.$$

Interestingly, for $\alpha = 2$ the algorithm collapses to the Box–Muller method [42] to generate normally distributed rvs.

As further discussed in Subsect. "The Generalized Hyperbolic Distribution", the mixing of normal distributions allows one to derive interesting distributions having support over the real line and exhibiting heavier tails than the Gaussian. While generally any distribution with support over the positive real line can be used as the mixing distribution for the variance, transformations of the positive alpha stable distribution are often used in financial modeling.

In this context the symmetric alpha stable distributions have a nice representation. In particular, if $X \sim S(\alpha^*, 0, c, 0)$ and A (independent of X) is an α/α^* positive alpha stable rv with scale parameter $\cos^{\alpha^*/\alpha}\left(\frac{\pi\alpha}{2\alpha^*}\right)$ then

$$Z = A^{1/\alpha^*}X \sim S(\alpha, 0, c, 0)\ .$$

For $\alpha^* = 2$ this property implies that every symmetric alpha stable distribution, i. e. an alpha stable distribution with $\beta = 0$, can be viewed as being conditionally normally distributed, i. e., it can be represented as a continuous variance normal mixture distribution.

Generally, the tail behavior of the resulting mixture distribution is completely determined by the (positive) tails of the variance mixing distribution. In the case of the positive alpha stable distribution this implies that the resulting symmetric stable distribution exhibits very heavy tails, in fact the second moment does not exist. As the literature is controversial on the adequacy of such heavy tails (see Sect. "Empirical Evidence About the Tails"), transformations of the positive alpha stable distribution are oftentimes considered to weight down the tails. The method of *exponential tilting* is very useful in this context. In a general setup the exponential tilting of a rv X with respect to a rv Y (defined on the same probability space) defines a new rv $\tilde{X}$, whose pdf can be written as

$$f_{\tilde{X}}(x;\theta) = f_X(x)\frac{\mathbb{E}\left[\exp(\theta Y)|X=x\right]}{\mathbb{E}\left[\exp(\theta Y)\right]}\,,$$

where the parameter θ determines the "degree of dampening". The exponential tilting of a rv X with respect to itself, known as *Esscher transformation*, is widely used in financial economics and mathematics, see e. g. [88]. In this case the resulting pdf is given by

$$\begin{aligned} f_{\tilde{X}}(x;\theta) &= \frac{\exp(\theta x)}{\mathbb{E}\left[\exp(\theta X)\right]}f_X(x) \\ &= \exp(\theta x - K(\theta))\,f_X(x)\,, \end{aligned} \qquad (13)$$

with $K(\cdot)$ denoting the cumulant generating function, $K(\theta) := \log\left(\mathbb{E}\left[\exp(\theta X)\right]\right)$.

The *tempered stable* (TS) laws are given by an application of the Esscher transform (13) to a positive alpha stable law. Note that the *Laplace transform* $\mathbb{E}[\exp(-tX)]$, $t \geq 0$, of a positive alpha stable rv is given by $\exp(-\delta\,(2t)^{\alpha})$, where $\delta = c^{\alpha}/(2^{\alpha}\cos(\frac{\pi\alpha}{2}))$. Thus, with $\theta = -(1/2)\gamma^{1/\alpha} \leq 0$, the pdf of the tempered stable law is given by

$$f_{\mathrm{TS}}(x;\alpha,\delta,\gamma) = \frac{\exp\left(-\frac{1}{2}\gamma^{1/\alpha}x\right)}{\mathbb{E}\left[\exp\left(-\frac{1}{2}\gamma^{1/\alpha}X\right)\right]} f_{\mathrm{S}}(x;\alpha,1,c(\delta,\alpha),0)$$
$$= \exp\left(\delta\gamma - \tfrac{1}{2}\gamma^{1/\alpha}x\right) f_{\mathrm{S}}(x;\alpha,1,c(\delta,\alpha),0)$$

with $0 < \alpha < 1$, $\delta > 0$, and $\gamma \geq 0$.

A generalization of the TS distribution was proposed by [22]. This class of *modified stable* (MS) laws can be obtained by applying the following transformation

$$f_{\mathrm{MS}}(x,\alpha,\lambda,\delta,\gamma) = c(\alpha,\lambda,\delta,\gamma)\,x^{\lambda+\alpha} f_{\mathrm{TS}}(x;\alpha,\delta,\gamma)\,, \tag{14}$$

with $\lambda \in \mathbb{R}$, $\gamma \vee (-\lambda) > 0$ and $c(\alpha,\lambda,\delta,\gamma)$ is a norming constant. For a more detailed analysis, we refer to [22]. Note that the terms “modified stable” or “tempered stable distribution” are not unique in the literature. Very often the so-called truncated Lévy flights/processes (see for example [56,130,157]) are also called TS processes (or corresponding distributions). These distributions are obtained by applying a smooth downweighting of the large jumps (in terms of the Lévy density).

The MS distribution is a quite flexible distribution defined over the positive real line and nests several important distributions. For instance, for $\alpha = 1/2$, the MS distribution reduces to the *generalized inverse Gaussian* (GIG) distribution, which is of main interest in Subsect. “The Generalized Hyperbolic Distribution”.

Note that in contrast to the unrestricted MS distribution, the pdf of the GIG distribution is available in closed form and can be straightforwardly obtained by applying the above transformation. In particular, for $\alpha = 1/2$, the positive alpha stable distribution is the Lévy distribution with closed form pdf given by

$$f_{\mathrm{S}}(x;1/2,1,c,0) = \sqrt{\frac{c}{2\pi}}\,\frac{\exp\left(-\frac{c}{2x}\right)}{x^{3/2}}\,.$$

Applying the Esscher transformation (13) with $\theta = -(1/2)\gamma^2$ yields the pdf of the inverse Gaussian (or Wald) distribution

$$f_{\mathrm{IG}}(x;\delta,\gamma) = \frac{\delta}{\sqrt{2\pi}}x^{-3/2}\exp\left(\delta\gamma - \left(\delta^2x^{-1}+\gamma^2x\right)/2\right),$$

where $\delta > 0$ and $\gamma \geq 0$. Applying the transformation (14) delivers the pdf of the GIG distribution,

$$f_{\mathrm{GIG}}(x;\lambda,\delta,\gamma) = \frac{(\gamma/\delta)^{\lambda}}{2K_{\lambda}(\delta\gamma)}x^{\lambda-1} \cdot \exp\left(-\left(\delta^2x^{-1}+\gamma^2x\right)/2\right), \tag{15}$$

with $K_{\lambda}(\cdot)$ being the *modified Bessel function of the third kind* and of order $\lambda \in \mathbb{R}$. Note that this function is oftentimes called the modified Bessel function of the second kind or Macdonald function. Nevertheless, one representation is given by

$$K_{\lambda}(x) = \frac{1}{2}\int_0^{\infty} y^{\lambda-1}\exp\left(-\frac{1}{2}x\left(y+y^{-1}\right)\right)\mathrm{d}y\,.$$

The parameters of the GIG distribution are restricted to satisfy the following conditions:

$$\begin{aligned} &\delta \geq 0 \text{ and } \gamma > 0 \quad \text{if } \lambda > 0\\ &\delta > 0 \text{ and } \gamma > 0 \quad \text{if } \lambda = 0\\ &\delta > 0 \text{ and } \gamma \geq 0 \quad \text{if } \lambda < 0\,. \end{aligned} \tag{16}$$

Importantly, in contrast to the positive alpha stable law, all moments exist and are given by

$$\mathbb{E}\left[X^r\right] = \left(\frac{\delta}{\gamma}\right)^r \frac{K_{\lambda+r}(\delta\gamma)}{K_{\lambda}(\delta\gamma)} \tag{17}$$

for all $r > 0$. For a very detailed analysis of the GIG distribution we refer to [117]. The GIG distribution nests several positive distributions as special cases and as limit distributions. Since all of these distributions belong to a special class of the generalized hyperbolic distribution, we proceed with a discussion of the latter, thus providing a broad framework for the discussion of many important distributions in finance.

The Generalized Hyperbolic Distribution

The mixing of normal distributions is well suited for financial modeling, as it allows construction of very flexible distributions. For example, the *normal variance-mean mixture*, given by

$$X = \mu + \beta V + \sqrt{V}Z\,, \tag{18}$$

with Z being normally distributed and V a positive random variable, generally exhibits heavier tails than the Gaussian distribution. Moreover, this mixture possesses interesting properties, for an overview see [26]. First, similarly to other mixtures, the normal variance-mean mixture is a conditional Gaussian distribution with conditioning on the volatility states, which is appealing when modeling financial returns. Second, if the mixing distribution, i. e. the distribution of V, is infinitely divisible, then X is likewise infinitely divisible. This implies that there exists a Lévy process with support over the whole real line, which is distributed at time $t = 1$ according to the law of X. As the theoretical properties of Lévy processes are well established in the literature (see, e. g., [24,194]), this result immediately suggests to formulate financial models in terms of the corresponding Lévy process.

Obviously, different choices for the distribution of V result in different distributions of X. However, based on the above property, an infinitely divisible distribution is most appealing. For the MS distributions discussed in Subsect. "Alpha Stable and Related Distributions", infinite divisibility is not yet established, although [22] strongly surmise so. However, for some special cases infinite divisibility has been shown to hold. The most popular is the GIG distribution yielding the *generalized hyperbolic* (GH) distribution for X. The latter distribution was introduced by [17] for modeling the distribution of the size of sand particles. The infinite divisibility of the GIG distribution was shown by [20].

To derive the GH distribution as a normal variance-mean mixture, let $V \sim \mathrm{GIG}(\lambda, \delta, \gamma)$ as in (15), with $\gamma = \sqrt{\alpha^2 - \beta^2}$, and $Z \sim \mathrm{N}(0, 1)$ independent of V. Applying (18) yields the GH distributed rv $X \sim \mathrm{GH}(\lambda, \alpha, \beta, \mu, \delta)$ with pdf

$$\begin{aligned} & f_{\mathrm{GH}}(x; \lambda, \alpha, \beta, \mu, \delta) \\ & = \frac{(\delta\gamma)^{\lambda}(\delta\alpha)^{1/2-\lambda}}{\sqrt{2\pi}\,\delta K_{\lambda}(\delta\gamma)} \left(1 + \frac{(x-\mu)^2}{\delta^2}\right)^{\lambda/2-1/4} \\ & \quad \cdot K_{\lambda-1/2}\left(\alpha\delta\sqrt{1+\frac{(x-\mu)^2}{\delta^2}}\right) \exp\left(\beta(x-\mu)\right) \end{aligned}$$

for $\mu \in \mathbb{R}$ and

$$\begin{aligned} &\delta \geq 0 \text{ and } |\beta| < \alpha \quad && \text{if } \lambda > 0 \\ &\delta > 0 \text{ and } |\beta| < \alpha \quad && \text{if } \lambda = 0 \\ &\delta > 0 \text{ and } |\beta| \leq \alpha \quad && \text{if } \lambda < 0\,, \end{aligned}$$

which are the induced parameter restrictions of the GIG distribution given in (16).

Note that, based on the mixture representation (18), the existing algorithms for generating GIG distributed rvs can be used to draw rvs from the GH distribution, see [12, 60].

For $|\beta + u| < \alpha$, the moment generating function of the GH distribution is given by

$$\begin{aligned} \mathbb{E}\left[\exp(uX)\right] &= \exp(\mu u)\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+u)^2}\right)^{\lambda/2} \\ & \quad \cdot \frac{K_{\lambda}\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}\,. \end{aligned} \tag{19}$$

As the moment generating function is infinitely differentiable in the neighborhood of zero, moments of all orders exist and have been derived in [25]. In particular, the mean and the variance of a GH rv X are given by

$$\begin{aligned} \mathbb{E}[X] &= \mu + \frac{\beta\delta}{\sqrt{\alpha^2-\beta^2}} \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \\ &= \mu + \beta\mathbb{E}[X_{\mathrm{GIG}}] \\ \mathbb{V}[X] &= \frac{\delta}{\sqrt{\alpha^2-\beta^2}} \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} + \frac{\beta^2\delta^2}{\alpha^2-\beta^2} \\ & \quad \cdot \left(\frac{K_{\lambda+2}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} - \frac{K^2_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K^2_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}\right) \\ &= \mathbb{E}[X_{\mathrm{GIG}}] + \beta^2\mathbb{V}[X_{\mathrm{GIG}}]\,, \end{aligned}$$

with $X_{\mathrm{GIG}} \sim \mathrm{GIG}(\lambda, \delta, \gamma)$ denoting a GIG distributed rv. Skewness and kurtosis can be derived in a similar way using the third and fourth derivative of the moment generating function (19). However, more information on the tail behavior is given by

$$f_{\mathrm{GH}}(x; \lambda, \alpha, \beta, \mu, \delta) \cong |x|^{\lambda-1} \exp\left((\mp\alpha + \beta)x\right)\,,$$

which shows that the GH distribution exhibits semi-heavy tails, see [22].

The moment generating function (19) also shows that the GH distribution is generally not closed under convolution. However, for $\lambda \in \{-1/2, 1/2\}$, the modified Bessel function of the third kind satisfies

$$K_{-\frac{1}{2}}(x) = K_{\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}} \exp(-x)\,,$$

yielding the following form of the moment generating function for $\lambda = -1/2$

$$\begin{aligned} & \mathbb{E}\left[\exp(uX) | \lambda = -1/2\right] \\ & \quad = \exp(\mu u) \frac{\exp\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{\exp\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}\,, \end{aligned}$$

which is obviously closed under convolution. This special class of the GH distribution is called normal inverse

Gaussian distribution and is discussed in more detail in Subsect. "The Normal Inverse Gaussian Distribution". The closedness under convolution is an attractive property of this distribution as it facilitates forecasting applications.

Another very popular distribution that is nested in the GH distribution is the hyperbolic distribution given by $\lambda = 1$ (see the discussion in Subsect. "The Hyperbolic Distribution"). Its popularity is primarily based on its pdf, which can (except for the norming constant) be expressed in terms of elementary functions allowing for a very fast numerical evaluation. However, given the increasing computer power and the slightly better performance in reproducing the unconditional return distribution, the normal inverse Gaussian distribution is now the most often used subclass.

Interestingly, further well-known distributions can be expressed as limiting cases of the GH distribution, when certain of its parameters approach their extreme values. To this end, the following alternative parametrization of the GH distribution turns out to be useful. Setting $\xi = 1/\sqrt{1+\delta\sqrt{\alpha^2-\beta^2}}$ and $\chi = \xi\beta/\alpha$ renders the two parameters invariant under location and scale transformations. This is an immediate result of the following property of the GH distribution. If $X \sim \mathrm{GH}(\lambda, \alpha, \beta, \mu, \delta)$, then

$$a + bX \sim \mathrm{GH}\left(\lambda, \frac{\alpha}{|b|}, \frac{\beta}{|b|}, a + b\mu, \delta\,|b|\right) .$$

Furthermore, the parameters are restricted by $0 \leq |\chi| < \xi < 1$, implying that they are located in the so-called *shape triangle* introduced by [27]. Figure 5 highlights the impact of the parameters in the GH distribution in terms of χ, ξ and λ. Obviously, χ controls the skewness and ξ the tailedness of the distribution. The impact of λ is not so clear-cut. The lower panels in Fig. 5 depict the pdfs for different values of λ whereby the first two moments and the values of χ and ξ are kept constant to show the "partial" influence.

As pointed out by [69], the limit distributions can be classified by the values of ξ and χ as well as by the values ϱ and ζ of a second location and scale invariant parametrization of the GH, given by $\varrho = \beta/\alpha$ and $\zeta = \delta\sqrt{\alpha^2-\beta^2}$, as follows:

- $\xi = 1$ and $-1 \leq \chi \leq 1$: The resulting limit distributions depend here on the values of λ. Note, that in order to reach the boundary either $\delta \to 0$ or $|\beta| \to \alpha$.
  - For $\lambda > 0$ and $|\beta| \to \alpha$ no limit distribution exists, but for $\delta \to 0$ the GH distribution converges to the distribution of a variance gamma rv (see Subsect. "The Variance Gamma Distribution"). However, note that $|\beta| < \alpha$ implies $|\chi| < \xi$ and so the limit distribution is not valid in the corners. For these cases, the limit distributions are given by $\xi = |\chi|$ and $0 < \xi \leq 1$, i. e. the next case.
  - For $\lambda = 0$ there exists no proper limit distribution.
  - For $\lambda < 0$ and $\delta \to 0$ no proper distribution exists but for $\beta \to \pm\alpha$ the limit distribution is given in [185] with pdf

$$\frac{2^{\lambda+1}\left(\delta^2+(x-\mu)^2\right)^{(\lambda-1/2)/2}}{\sqrt{2\pi}\,\Gamma(-\lambda)\,\delta^{2\lambda}\alpha^{\lambda-1/2}} \cdot K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right) \cdot \exp\left(\pm\alpha\,(x-\mu)\right) , \qquad (20)$$

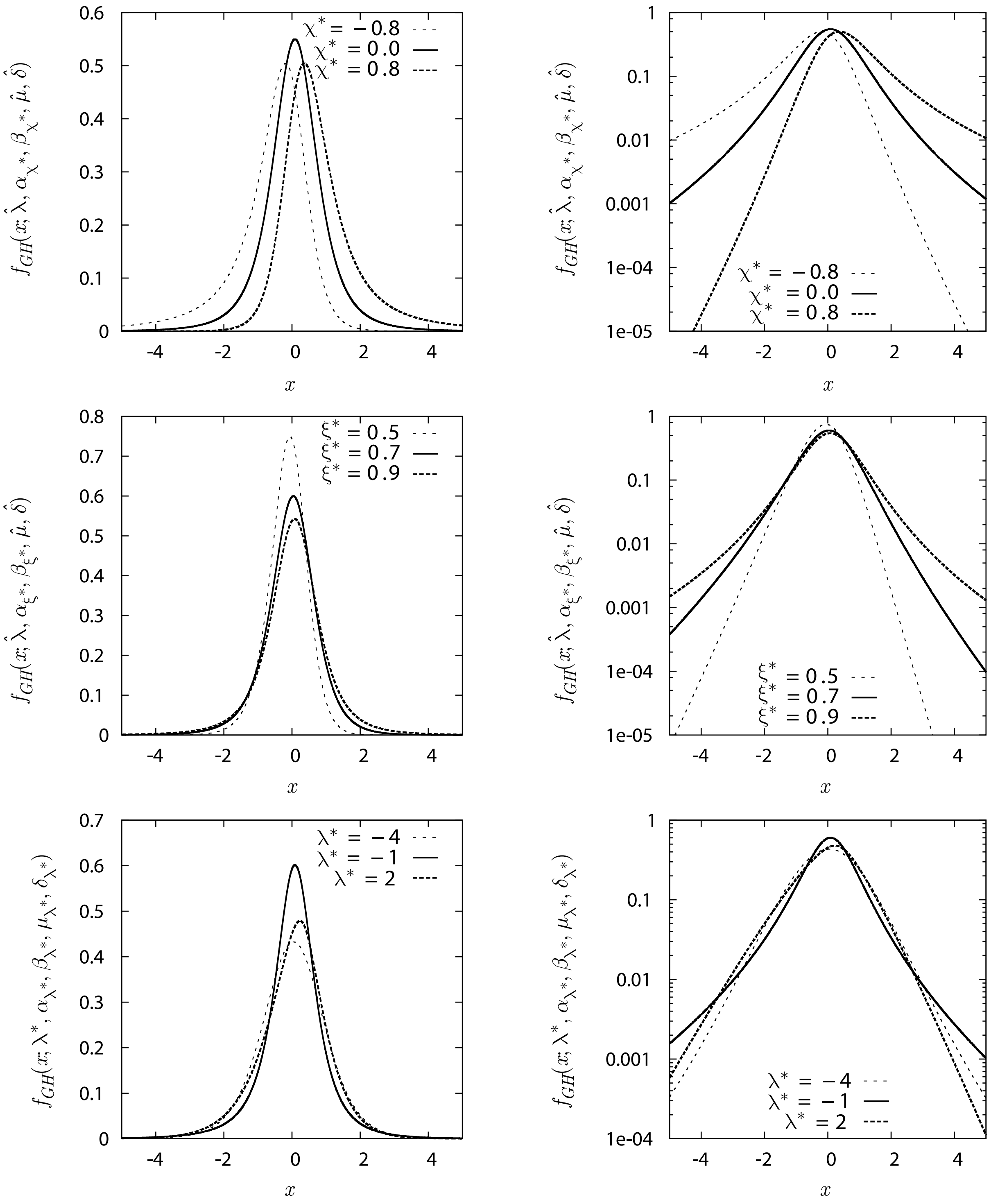
  which is the limit distribution of the corners, since $\beta = \pm\alpha$ is equivalent to $\chi = \pm\xi$. This distribution was recently called the GH skew t distribution by [1]. Assuming additionally that $\alpha \to 0$ and $\beta = \varrho\alpha \to 0$ with $\varrho \in (-1, 1)$ yields the limit distribution in between

$$\frac{\Gamma(-\lambda+1/2)}{\sqrt{\pi\delta^2}\,\Gamma(-\lambda)}\left(1+\frac{(x-\mu)^2}{\delta^2}\right)^{\lambda-1/2} ,$$

  which is the scaled and shifted Student's t distribution with -2λ degrees of freedom, expectation μ and variance $4\lambda^2\nu/(\nu-2)$, for more details see Subsect. "The Student t Distribution".
- $\xi = |\chi|$ and $0 < \xi \leq 1$: Except for the upper corner the limit distribution of the right boundary can be derived for

$$\beta = \alpha - \frac{\phi}{2}\,; \quad \alpha \to \infty\,; \quad \delta \to 0\,; \quad \alpha\delta^2 \to \tau$$

  with $\phi > 0$ and is given by the μ-shifted GIG distribution $\mathrm{GIG}\left(\lambda, \sqrt{\tau}, \sqrt{\phi}\right)$. The distribution for the left boundary is the same distribution but mirrored at $x = 0$. Note that the limit behavior does not depend on λ. However, to obtain the limit distributions for the left and right upper corners we have to distinguish for different values of λ. Recall that for the regime $\xi = 1$ and $-1 \leq \chi \leq 1$ the derivation was not possible.
  - For $\lambda > 0$ the limit distribution is a gamma distribution.
  - For $\lambda = 0$ no limit distribution exists.
  - For $\lambda < 0$ the limit distribution is the reciprocal gamma distribution.
- $\xi = \chi = 0$: This is the case for $\alpha \to \infty$ or $\delta \to \infty$. If only $\alpha \to \infty$ then the limit distribution is the Dirac measure concentrated in μ. If in addition $\delta \to \infty$ and $\delta/\alpha \to \sigma^2$ then the limit distribution is a normal distribution with mean $\mu + \beta\sigma^2$ and variance σ^2.



Financial Economics, Fat-Tailed Distributions, Figure 5

Density function (pdf) of the GH distribution for different parameter vectors. The *right panel* plots the log-densities to better visualize the tail behavior. The *upper and middle section* present the pdf for different values of χ and ξ. Note that these correspond to different values of α and β. The *lower panel* highlights the influence of λ if the first two moments, as well as χ and ξ, are held fixed. This implies that α, β, μ and δ have to be adjusted accordingly

As pointed out by [185] applying the unrestricted GH distribution to financial data results in a very flat likelihood function especially for λ. This characteristic is illustrated in Fig. 6, which plots the maximum of the log likelihood for different values of λ using our sample of the S&P500 index returns. This implies that the estimate of λ is generally associated with a high standard deviation. As a consequence, rather than using the GH distribution directly, the finance literature primarily predetermines the value of λ, resulting in specific subclasses of the GH distribution, which are discussed in the sequel. However, it is still interesting to derive the general results in terms of the GH distribution (or the corresponding Lévy process) directly and to restrict only the empirical application to a subclass. For example [191] derived a diffusion process with GH marginal distribution, which is a generalization of the result of [33], who proposed a diffusion process with hyperbolic marginal distribution.

The Hyperbolic Distribution Recall, that the *hyperbolic* (HYP) distribution can be obtained as a special case of the GH distribution by setting $\lambda = 1$. Thus, all properties of the GH law can be applied to the HYP case. For instance the pdf of the HYP distribution is straightforwardly given by (19) setting $\lambda = 1$

$$\begin{aligned} f_{\mathrm{H}}(x;\alpha,\beta,\mu,\delta) &:= f_{\mathrm{GH}}(x;1,\alpha,\beta,\mu,\delta) \\ &= \frac{\sqrt{\alpha^2-\beta^2}}{2\alpha\delta K_1\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \\ &\quad \cdot \exp\left(-\alpha\sqrt{\delta 2+(x-\mu)2}\right. \\ &\quad \left.+\beta(x-\mu)\right)\,, \end{aligned} \tag{21}$$

where $0 \leq |\beta| < \alpha$ are the shape parameter and $\mu \in \mathbb{R}$ and $\delta > 0$ are the location and scale parameter, respectively.

The distribution was applied to stock return data by [70,71,132] while [33] derived a diffusion model with marginal distribution belonging to the class of HYP distributions.

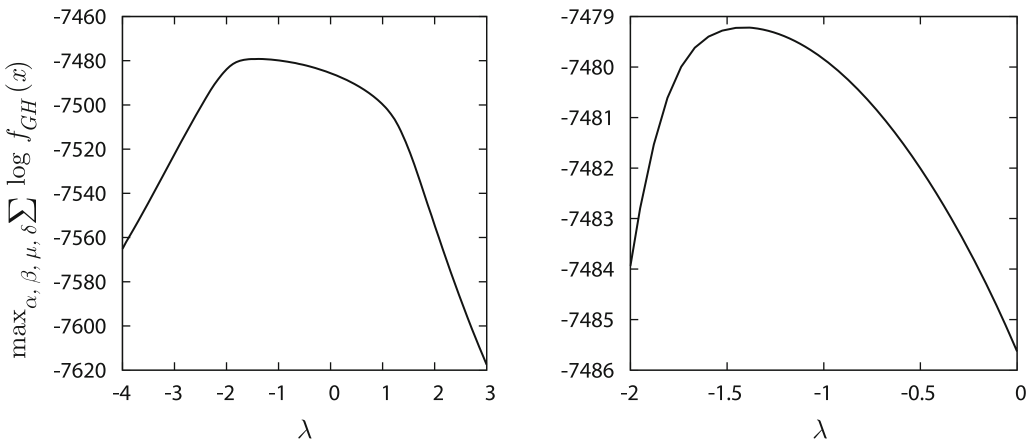
The Normal Inverse Gaussian Distribution The *normal inverse Gaussian* (NIG) distribution is given by the GH distribution with $\lambda = -1/2$ and has the following pdf

$$\begin{aligned} f_{\mathrm{NIG}}(x;\alpha,\beta,\mu,\delta) &:= f_{\mathrm{GH}}\left(x;-\frac{1}{2},\alpha,\beta,\mu,\delta\right) \\ &= \frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\pi\sqrt{\delta^2+(x-\mu)^2}} \\ &\quad \cdot \exp\left(\delta\gamma+\beta(x-\mu)\right) \end{aligned} \tag{22}$$

with $0 < |\beta| \leq \alpha$, $\delta > 0$ and $\mu \in \mathbb{R}$. The moments of a NIG distributed rv can be obtained from the moment generating function of the GH distribution (19) and are given by

$$\mathbb{E}[X] = \mu + \frac{\delta\beta}{\sqrt{\alpha^2-\beta^2}} \quad \text{and} \quad \mathbb{V}[X] = \frac{\delta\alpha^2}{\sqrt{\alpha^2-\beta^2}^3}$$
$$\mathbb{S}[X] = 3\frac{\beta}{\alpha\sqrt{\delta\sqrt{\alpha^2-\beta^2}}} \quad \text{and} \quad \mathbb{K}[X] = 3\frac{\alpha^2+4\beta^2}{\delta\alpha^2\sqrt{\alpha^2-\beta^2}}\,.$$

This distribution was heavily applied in financial economics for modeling the unconditional as well as the conditional return distribution, see e. g. [18,21,185]; as well



Financial Economics, Fat-Tailed Distributions, Figure 6
Partially maximized log likelihood, estimated maximum log likelihood values of the GH distribution for different values of λ

as [10,19,114], respectively. Recently, [57] used the NIG distribution for modeling realized variance and found improved forecast performance relative to a Gaussian model. A more realistic modeling of the distributional properties is not only important for risk management or forecasting, but also for statistical inference. For example the efficient method of moments, proposed by [87] requires the availability of a highly accurate auxiliary model, which provide the objective function to estimate a more structural model. Recently, [39] provided such an auxiliary model, which uses the NIG distribution and realized variance measures.

Recall that for $\lambda = -1/2$ the mixing distribution is the inverse Gaussian distribution, which facilitates the generation of rvs. Hence, rvs with NIG distribution can be generated in the following way:

1. Draw a chi-square distributed rv C with one degree of freedom and a uniformly distributed rv over the interval $(0, 1)$ U
2. Compute

$$X_1 = \frac{\delta}{\gamma} + \frac{1}{2\delta\gamma}\left(\frac{\delta C}{\gamma} - \sqrt{4\delta^3 C/\gamma + \delta^2 C^2/\gamma^2}\right)$$

3. If $U < \delta/(\gamma(\delta/\gamma + X_1))$ return X_1 else return $\delta^2/(\gamma^2 X_1)$.

As pointed out by [187] the main difference between the HYP and NIG distribution: "Hyperbolic log densities, being hyperbolas, are strictly concave everywhere. Therefore they cannot form any sharp tips near $x = 0$ without loosing too much mass in the tails … In contrast, NIG log densities are concave only in an interval around $x = 0$, and convex in the tails." Moreover, [19] concludes, "It is, moreover, rather typical that asset returns exhibit tail behavior that is somewhat heavier than log linear, and this further strengthens the case for the NIG in the financial context".

The Student *t* Distribution Next to the alpha stable distribution *Student's t* (*t* thereafter) distribution has the longest history in financial economics. One reason is that although the non-normality of asset returns is widely accepted, there still exists some discussion on the exact tail behavior. While the alpha stable distribution implies extremely slowly decreasing tails for $\alpha \neq 2$, the t distribution exhibits power tails and existing moments up to (and excluding) ν. As such, the t distribution might be regarded as the strongest competitor to the alpha stable distribution, shedding also more light on the empirical tail behavior of returns. The pdf for the scaled and shifted t distribution is given by

$$f_t(x;\nu,\mu,\sigma) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)\,\sigma} \cdot \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \tag{23}$$

for $\nu > 0$, $\sigma > 0$ and $\mu \in \mathbb{R}$. For $\mu = 0$ and $\sigma = 1$ the well-known standard t distribution is obtained. The shifted and scaled t distribution can also be interpreted as a mean-variance mixture (18) with a reciprocal gamma distribution as a mixing distribution. The mean, variance, and kurtosis (3) are given by μ, $\sigma^2\nu/(\nu-2)$, and $3(\nu-2)/(\nu-4)$, provided that $\nu > 1$, $\nu > 2$, and $\nu > 4$, respectively. The tail behavior is

$$f_t(x,\nu,\mu,\sigma) \cong c x^{-\nu-1}\,.$$

The t distribution is one of the standard non-normal distributions in financial economics, see e. g. [36,38,184]. However, as the unconditional return distribution may exhibit skewness, a skewed version of the t distribution might be more adequate in some cases. In fact, several skewed t distributions were proposed in the literature, for a short overview see [1]. The following special form of the pdf was considered in [81,102]

$$\begin{aligned} f_{t,\mathrm{FS}}(x;\nu,\mu,\sigma,\beta) &= \frac{2\beta}{\beta^2+1}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma} \\ &\quad \cdot \left(1+\tfrac{1}{\nu}\left(\tfrac{x-\mu}{\sigma}\right)^2\left(\tfrac{1}{\beta^2}\mathcal{I}(X\geq\mu)+\beta^2\mathcal{I}(x<\mu)\right)\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

with $\beta > 0$. For $\beta = 1$ the pdf reduces to the pdf of the usual symmetric scaled and shifted t distribution. Another skewed t distribution was proposed by [116] with pdf

$$\begin{aligned} f_{t,\mathrm{JF}}(x,\nu,\mu,\sigma,\beta) &= \frac{\Gamma(\nu+\beta)\,2^{1-\nu-\beta}}{\Gamma(\nu/2)\,\Gamma(\nu/2+\beta)\sqrt{\nu+\beta}\sigma} \\ &\quad \cdot \left(1+\frac{\frac{x-\mu}{\sigma}}{\sqrt{\nu+\beta+\left(\frac{x-\mu}{\sigma}\right)^2}}\right)^{(\nu+1)/2} \\ &\quad \cdot \left(1-\frac{\frac{x-\mu}{\sigma}}{\sqrt{\nu+\beta+\left(\frac{x-\mu}{\sigma}\right)^2}}\right)^{\beta+(\nu+1)/2} \end{aligned}$$

for $\beta > -\nu/2$. Again, the usual t distribution can be obtained as a special case for $\beta = 0$. A skewed t distribution in terms of the pdf and cdf of the standard t distribution $f_t(x;\nu,0,1)$ and $F_t(x;\nu,0,1)$ is given by [13,43]

$$
\begin{aligned}
&f_{t,\mathrm{AC}}(x;\nu,\mu,\sigma,\beta)\\
&\quad=\frac{2}{\sigma}f_t\left(\frac{x-\mu}{\sigma},\nu,0,1\right)\\
&\quad\cdot F_t\left(\beta\left(\frac{x-\mu}{\sigma}\right)\sqrt{\frac{\nu+1}{\nu+\left(\frac{x-\mu}{\sigma}\right)^2}},\nu+1,0,1\right)
\end{aligned}
$$

for $\beta \in \mathbb{R}$.

Alternatively, a skewed t distribution can also be obtained as a limit distribution of the GH distribution. Recall that for $\lambda < 0$ and $\beta \to \alpha$ the limit distribution is given by (20) as

$$
\begin{aligned}
&f_{t,\mathrm{GH}}(x;\lambda,\mu,\delta,\alpha)\\
&\quad=\frac{2^{\lambda+1}\left(\delta^2+(x-\mu)^2\right)^{(\lambda-1/2)/2}}{\sqrt{2\pi}\,\Gamma(-\lambda)\,\delta^{2\lambda}\alpha^{\lambda-1/2}}\\
&\quad\cdot K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)\cdot\exp\left(\alpha(x-\mu)\right)\,.
\end{aligned}
$$

for $\alpha \in \mathbb{R}$. The symmetric t distribution is obtained for $\alpha \to 0$. The distribution was introduced by [185] and a more detailed examination was recently given in [1].

The Variance Gamma Distribution The *variance gamma* (VG) distribution can be obtained as a mean-variance mixture with gamma mixing distribution. Note that the gamma distribution is obtained in the limit from the GIG distribution for $\lambda > 0$ and $\delta \to 0$. The pdf of the VG distribution is given by

$$
\begin{aligned}
f_{\mathrm{VG}}(x;\mu,\alpha,\beta,\lambda) &:= \lim_{\delta\to 0} f_{\mathrm{GH}}(x;\lambda,\alpha,\beta,\delta,\mu)\\
&=\frac{\gamma^{2\lambda}|x-\mu|^{\lambda-1/2}K_{\lambda-1/2}(\alpha|x-\mu|)}{\sqrt{\pi}\,\Gamma(\lambda)\,(2\alpha)^{\lambda-1/2}}\exp\beta(x-\mu)\,.
\end{aligned}
\tag{24}
$$

Note, the usual parameterization of the VG distribution

$$
\begin{aligned}
&f^*_{\mathrm{VG}}\left(x;\sigma^*,\theta^*,\nu^*,\mu^*\right)\\
&\quad=\frac{2\exp\left(\theta^*(x-\mu^*)/\sigma^{*2}\right)}{\nu^{*1/\nu^*}\sqrt{2\pi\sigma^{*2}}\,\Gamma(1/\nu^*)}\left(\frac{(x-\mu^*)^2}{2\sigma^{*2}/\nu^*+\theta^{*2}}\right)^{\frac{1}{2\nu^*}-\frac{1}{4}}\\
&\quad\cdot K_{\frac{1}{\nu^*}-\frac{1}{2}}\left(\frac{\sqrt{(x-\mu^*)^2\left(2\sigma^{*2}/\nu^*+\theta^{*2}\right)}}{\sigma^{*2}}\right)
\end{aligned}
$$

is different from the one used here, however the parameters can be transformed between these representations in the following way

$$
\begin{aligned}
\sigma^* &= \sqrt{\frac{2\lambda}{\alpha^2-\beta^2}}\,; & \theta^* &= \frac{2\beta\lambda}{\alpha^2-\beta^2}\,;\\
\nu^* &= \frac{1}{\lambda}\,; & \mu^* &= \mu\,.
\end{aligned}
$$

This distribution was introduced by [149,150,151]. For $\lambda = 1$ (the HYP case) we obtain a skewed, shifted and scaled Laplace distribution with pdf

$$
\begin{aligned}
&f_{\mathrm{L}}(x;\alpha,\beta,\mu)\\
&\quad:=\lim_{\delta\to 0} f_{\mathrm{GH}}(x;1,\alpha,\beta,\delta,\mu)\\
&\quad=\frac{\alpha^2-\beta^2}{2\alpha}\exp\left(-\alpha|x-\mu|+\beta(x-\mu)\right)\,.
\end{aligned}
$$

A generalization of the VG distribution to the so-called CGMY distribution was proposed by [48].

The Cauchy Distribution Setting $\lambda = -1/2$, $\beta \to 0$ and $\alpha \to 0$ the GH distribution converges to the *Cauchy* distribution with parameters μ and δ. Since the Cauchy distribution belongs to the class of symmetric alpha stable ($\alpha = 1$) and symmetric t distributions ($\nu = 1$) we refer to Subsect. "Alpha Stable and Related Distributions" and "The Student t Distribution" for a more detailed discussion.

The Normal Distribution For $\alpha \to \infty$, $\beta = 0$ and $\delta = 2\sigma^2$ the GH distribution converges to the *normal* distribution with mean μ and variance σ^2.

Finite Mixtures of Normal Distributions

The density of a (finite) mixture of k normal distributions is given by a linear combination of k Gaussian *component densities*, i. e.,

$$
\begin{aligned}
f_{\mathrm{NM}}(x;\theta) &= \sum_{j=1}^{k}\lambda_j\phi\left(x;\mu_j,\sigma_j^2\right)\,,\\
\phi\left(x;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,,
\end{aligned}
\tag{25}
$$

where $\theta = (\lambda_1,\dots,\lambda_{k-1},\mu_1,\dots,\mu_k,\sigma_1^2,\dots,\sigma_k^2)$, $\lambda_k = 1-\sum_{j=1}^{k-1}\lambda_j$, $\lambda_j > 0$, $\mu_j \in \mathbb{R}$, $\sigma_j^2 > 0$, $j = 1,\dots,k$, and $(\mu_i,\sigma_i^2) \neq (\mu_j,\sigma_j^2)$ for $i \neq j$. In (25), the λ_j, μ_j, and σ_j^2 are called the *mixing weights*, *component means*, and *component variances*, respectively.

Finite mixtures of normal distributions have been applied as early as 1886 in [174] to model leptokurtic phenomena in astrophysics. A voluminous literature exists, see [165] for an overview. In our discussion, we shall focus on a few aspects relevant for applications in finance. In this context, (25) arises naturally when the component densities are interpreted as different *market regimes*. For example, in a two-component mixture ($k = 2$), the first component, with a relatively high mean and small variance, may be interpreted as the bull market regime, occurring with probability λ_1, whereas the second regime, with a lower expected return and a greater variance, represents the bear market. This (typical) pattern emerges for the S&P500 returns, see Table 1. Clearly (25) can be generalized to accommodate non-normal component densities; e. g., [104] model stock returns using mixtures of generalized error distributions of the form (40). However, it may be argued that in this way much of the original appeal of (25), i. e., within-regime normality along with CLT arguments, is lost.

The moments of (25) can be inferred from those of the normal distribution, with mean and variance given by

$$\begin{aligned} \mathbb{E}\,[X] &= \sum_{j=1}^{k} \lambda_j \mu_j\,, \text{ and} \\ \mathbb{V}\,[X] &= \sum_{j=1}^{k} \lambda_j \left(\sigma_j^2 + \mu_j^2\right) - \left(\sum_{j=1}^{k} \lambda_j \mu_j\right)^2 , \end{aligned} \tag{26}$$

respectively. The class of finite normal mixtures is very flexible in modeling the leptokurtosis and, if existent, skewness of financial data. To illustrate the first property, consider the *scale normal mixture*, where, in (25), $\mu_1 = \mu_2 = \cdots = \mu_k := \mu$. In fact, when applied to financial return data, it is often found that the market regimes differ mainly in their variances, while the component means are rather close in value, and often their differences are not significant statistically. This reflects the observation that excess kurtosis is a much more pronounced (and ubiquitous) property of asset returns than skewness. In the scale mixture case, the density is symmetric, but with higher peaks and thicker tails than the normal with the same mean and variance. To see this, note that $\sum_j(\lambda_j/\sigma_j) > (\sum_j \lambda_j\sigma_j^2)^{-1/2} \Leftrightarrow (\sum_j \lambda_j\sigma_j^2)^{1/2} > [\sum_j(\lambda_j/\sigma_j)]^{-1}$. But $(\sum_j \lambda_j\sigma_j^2)^{1/2} > \sum_j \lambda_j\sigma_j > [\sum_j(\lambda_j/\sigma_j)]^{-1}$ by Jensen's and the arithmetic-harmonic mean inequality, respectively. This shows $f_{\mathrm{NM}}(\mu;\theta) > \phi(\mu;\mu,\sum_j \lambda_j\sigma_j^2)$, i. e., peakedness. Tailedness follows from the observation that the difference between the mixture and the mean-variance equivalent normal density is asymptotically dominated by the component with the greatest variance. Moreover, the densities of the scale mixture and the mean-variance equivalent Gaussian intersect exactly two times on both sides of the mean, so that the scale mixture satisfies the density crossing condition in Finucan's theorem mentioned in Sect. "Definition of the Subject" and observed in the center panel of Fig. 1. This follows from the fact that, if $a_1, \ldots, a_n$ and $\gamma_1 < \cdots < \gamma_n$ are real con-

Financial Economics, Fat-Tailed Distributions, Table 1
Maximum-likelihood parameter estimates of the iid model

Distri-bution	Parameters					Loglik
GH	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	−7479.2
	−1.422	0.087	0.322	−0.046	1.152	
	(0.351)	(0.018)	(0.222)	(0.022)	(0.139)	
t_{GH}	$\hat{\mu}$	$\hat{\delta}$	$\hat{\lambda}$	$\hat{\alpha}$		−7479.7
	0.084	1.271	3.445	−0.041		
	(0.018)	(0.052)	(0.181)	(0.021)		
t_{JF}	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$		−7480.0
	3.348	0.098	0.684	0.091		
	(0.179)	(0.025)	(0.012)	(0.049)		
t_{AC}	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$		−7480.1
	3.433	0.130	0.687	−0.123		
	(0.180)	(0.042)	(0.013)	(0.068)		
t_{FS}	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$		−7480.3
	3.432	0.085	0.684	0.972		
	(0.180)	(0.020)	(0.012)	(0.017)		
Symmetric t	$\hat{\nu}$	$\hat{\mu}$	$\hat{\sigma}$			−7481.7
	3.424	0.056	0.684			
	(0.179)	(0.011)	(0.012)			
NIG	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$		−7482.0
	0.088	0.784	−0.048	0.805		
	(0.018)	(0.043)	(0.022)	(0.028)		
HYP	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$		−7499.5
	0.090	1.466	−0.053	0.176		
	(0.018)	(0.028)	(0.023)	(0.043)		
VG	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\lambda}$		−7504.2
	0.092	1.504	−0.054	1.115		
	(0.013)	(0.048)	(0.019)	(0.054)		
Alpha stable	$\hat{\alpha}$	$\hat{\beta}$	$\hat{c}$	$\hat{\tau}$		−7522.5
	1.657	−0.094	0.555	0.036		
	(0.024)	(0.049)	(0.008)	(0.015)		
Finite mixture ($k = 2$)	$\hat{\lambda}_1$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	−7580.8
	0.872	0.063	−0.132	0.544	4.978	
	(0.018)	(0.012)	(0.096)	(0.027)	(0.530)	
Cauchy	$\hat{\mu}$	$\hat{\sigma}$				−7956.6
	0.060	0.469				
	(0.010)	(0.008)				
Normal	$\hat{\mu}$	$\hat{\sigma}$				−8168.9
	0.039	1.054				
	(0.014)	(0.010)				

Shown are maximum likelihood estimates for iid models with different assumptions about the distribution of the innovations. Standard errors are given in parentheses. "Loglik" is the value of the maximized log likelihood function.

stants, and N is the number of real zeros of the function $\varphi(x) = \sum_i a_i e^{\gamma_i x}$, then $W - N$ is a non-negative even integer, where W is the number of sign changes in the sequence $a_1, \ldots, a_n$ [183]. Skewness can be incorporated into the model when the component means are allowed to differ. For example, if, in the two-component mixture, the high-variance component has both a smaller mean and mixing weight, then the distribution will be skewed to the left.

Because of their flexibility and the aforementioned economic interpretation, finite normal mixtures have been frequently used to model the unconditional distribution of asset returns [40,44,129,179], and they have become rather popular since the publication of Hamilton's [101] paper on Markov-switching processes, where the mixing weights are assumed to be time-varying according to a k-state Markov chain; see, e. g., [200] for an early contribution in this direction.

However, although a finite mixture of normals is a rather flexible model, its tails decay eventually in a Gaussian manner, and therefore, according to the discussion in Sect. "Empirical Evidence About the Tails", it may often not be appropriate to model returns at higher frequencies unconditionally. Nevertheless, when incorporated into a GARCH structure (see Sect. "Volatility Clustering and Fat Tails"), it provides a both useful and intuitively appealing framework for modeling the *conditional* distribution of asset returns, as in [5,96,97]. These papers also provide a discussion of alternative interpretations of the mixture model (25), as well as an overview over the extensive literature.

Empirical Comparison

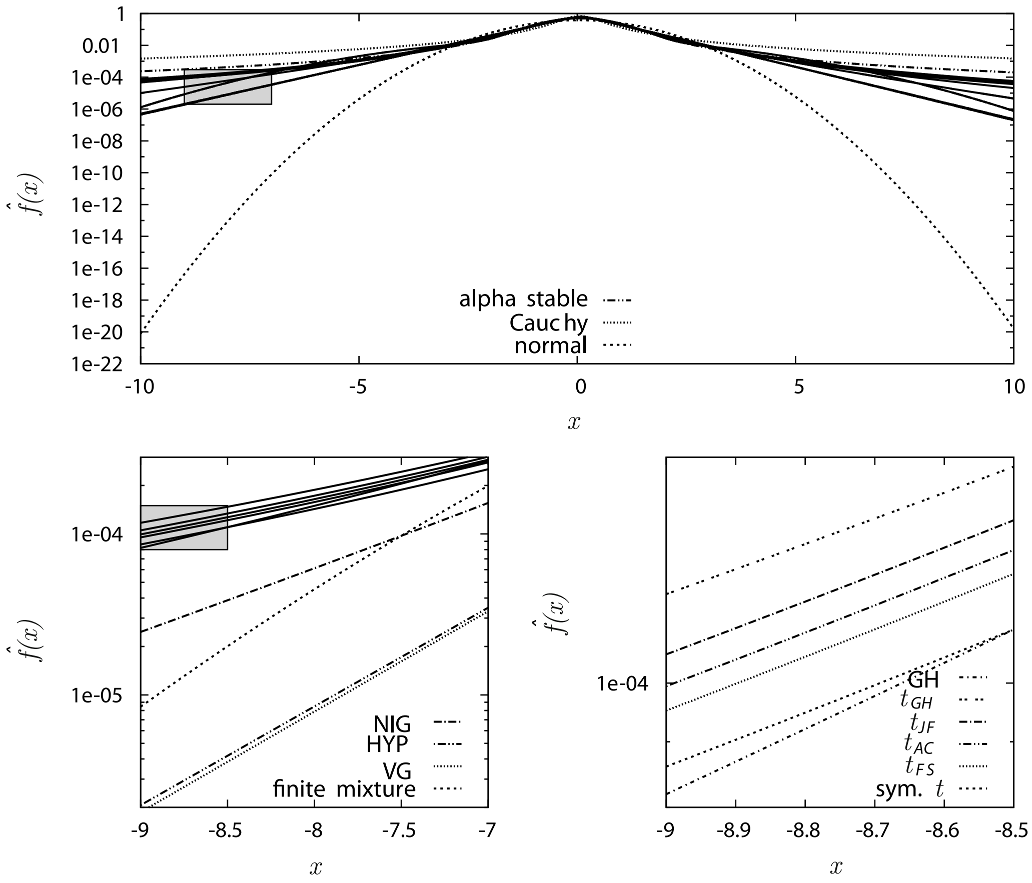
In the following we empirically illustrate the adequacy of the various distributions discussed in the previous sections for modeling the unconditional return distribution. Table 1 presents the estimation results for the S&P500 index assuming iid returns. The log likelihood values clearly indicate the inadequacy of the normal, Cauchy and stable distributions. This is also highlighted in the upper panel of Fig. 7, which clearly shows that the tails of the Cauchy and stable distributions are too heavy, whereas those of the normal distribution are too weak. To distinguish the other distributions in more detail, the lower left panel is an enlarged display of the shadowed box in the upper panel. It illustrates nicely that the two component mixture, VG and HYP distribution exhibit semiheavy tails, which are probably a little bit to weak for an adequate modeling as is indicated by the log likelihood values. Similarly, the two-component finite normal mixture, although much better than the normal, cannot keep up with most of the other models, presumably due to its essentially Gaussian tails. Although the pdf of the NIG distribution lies somewhere in between the pdfs of the HYP and the different t distributions, the log likelihood value clearly indicates that this distribution is in a statistical sense importantly closer to the t distributions. A further distinction between the other distributions including all kinds of t distributions and the GH distribution is nearly impossible, as can be seen from the lower right plot, which is an enlarged display of the lower left panel. The log likelihood values also do not allow for a clear distinction. Note also that the symmetric t distribution performs unexpectedly well. In particular, its log likelihood is almost indistinguishable from those of the skewed versions. Also note that, for all t distributions, the estimated tail index, ν, is close to 3.5, which is in accordance with the results from semiparametric tail estimation in Sect. "Empirical Evidence About the Tails".

The ranking of the distributions in terms of the log likelihood depends of course heavily on the dataset, and different return series may imply different rankings. However, Table 1 also highlights some less data-dependent results, which are more or less accepted in the literature, e. g., the tails of the Cauchy and stable distributions are too heavy, and those of the HYP and VG are too light for the unconditional distribution. This needs of course no longer be valid in a different modeling setup. Especially in a GARCH framework the conditional distribution don't need to imply such heavy tails because the model itself imposes fatter tails.

In Sect. "Application to Value-at-Risk", the comparison of the models will be continued on the basis of their ability the measure the Value-at-Risk, an important concept in risk management.

Volatility Clustering and Fat Tails

It has long been known that the returns of most financial assets, although close to being unpredictable, exhibit significant dependencies in measures of volatility, such as absolute or squared returns. Moreover, the empirical results based on the recent availability of more precise volatility measures, such as the *realized volatility*, which is defined as the sum over the squared intradaily high-frequency returns (see, e. g., [7] and [23]), also point towards the same direction. In particular, the realized volatility has been found to exhibit strong persistence in its autocorrelation function, which shows a hyperbolic decay indicating the presence of long memory in the volatility process. In fact, this finding as well as other stylized features of the realized



Financial Economics, Fat-Tailed Distributions, Figure 7
Plot of the estimated pdfs of the different return distributions assuming iid returns

volatility have been observed across different data sets and markets and are therefore by now widely acknowledged and established in the literature. For a more detailed and originating discussion on the stylized facts of the high-frequency based volatility measures for stock returns and exchange returns we refer to [8,9], respectively.

The observed dependence of time-varying pattern of the volatility is usually referred to as *volatility clustering*. It is also apparent in the top panel of Fig. 1 and was already observed by Mandelbrot [155], who noted that "large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes". It is now well understood that volatility clustering can explain at least part of the fat-tailedness of the unconditional return distribution, even if the *conditional* distribution is Gaussian. This is also supported by the recent observation that if the returns are scaled by the realized volatility then the distribution of the resulting series is approximately Gaussian (see [9] and [8]). To illustrate, consider a time series $\{\epsilon_t\}$ of the form

$$\epsilon_t = \eta_t \sigma_t \ , \tag{27}$$

where $\{\eta_t\}$ is an iid sequence with mean zero and unit variance, with η_t being independent of σ_t, so that σ_t^2 is the conditional variance of ϵ_t. With respect to the kurtosis measure $\mathbb{K}$ in (3), it has been observed by [108], and earlier by [31] in a different context, that, as long as σ_t^2 is not constant, Jensen's inequality implies $\mathbb{E}[\epsilon_t^4] = \mathbb{E}[\eta_t^4]\mathbb{E}[\sigma_t^4] > \mathbb{E}[\eta_t^4]\mathbb{E}^2[\sigma_t^2]$, so that the kurtosis of the unconditional distribution exceeds that of the innovation process. Clearly, $\mathbb{K}$ provides only limited information about the actual shape of the distribution, and more meaningful results can be obtained by specifying the dynamics of the conditional variance, σ_t^2. A general useful result [167] for analyzing the tail behavior of processes such as (27) is that, if ξ_t and σ_t are independent non-negative random variables with σ_t regularly varying, i. e., $P(\sigma_t > x) = L(x)x^{-\alpha}$ for some slowly varying L, and $\mathbb{E}[\xi_t^{\alpha+\delta}] < \infty$ for some $\delta > 0$, then

$\xi_t \sigma_t$ is likewise regularly varying with tail index α, namely,

$$P(\xi_t \sigma_t > x) \cong \mathbb{E}\left[\xi^\alpha\right] P(\sigma_t > x) \quad \text{as } x \to \infty \,. \tag{28}$$

Arguably the most popular model for the evolution of σ_t^2 in (27) is the generalized autoregressive conditional heteroskedasticity process of orders p and q, or GARCH(p, q), as introduced by [37,73], which specifies the conditional variance as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \,. \tag{29}$$

The case $p = 0$ in (29) is referred to as an ARCH(q) process, which is the specification considered in [73]. To make sure that the conditional variance remains positive for all t, appropriate restrictions have to be imposed on the parameters in (29), i. e., α_i, $i = 0, \ldots, q$, and β_i, $i = 1, \ldots, p$. It is clearly sufficient to assume that α_0 is positive and all the other parameters are non-negative, as in [37], but these conditions can be relaxed substantially if $p, q > 0$ and $p + q > 2$ [173].

(27) and (29) is covariance stationary iff

$$P(z) = z^m - \sum_{i=1}^{m} (\alpha_i + \beta_i) z^{m-i} = 0 \Rightarrow |z| < 1 \,, \tag{30}$$

where $m = \max\{p, q\}$, and $\alpha_i = 0$ for $i > q$, and $\beta_i = 0$ for $i < p$, which boils down to $\sum_i \alpha_i + \sum_i \beta_i < 1$ in case the non-negativity restrictions of [37] are imposed. The situation $\sum_i \alpha_i + \sum_i \beta_i = 1$ is referred to as an integrated GARCH (IGARCH) model, and in applications it is often found that the sum is just below unity. This indicates a high degree of volatility persistence, but the interpretation of this phenomenon is not so clear-cut [166]. If (30) holds, the unconditional variance of the process defined by (27) and (29) is given by

$$\mathbb{E}\left[\epsilon_t^2\right] = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{i=1}^{p} \beta_i} \,. \tag{31}$$

In practice, the GARCH(1,1) specification is of particular importance, and it will be the focus of our discussion too, i. e., we shall concentrate on the model (27) with

$$\sigma_t^2 = \alpha_0 + \left(\alpha_1 \eta_{t-1}^2 + \beta_1\right) \sigma_{t-1}^2 \,, \qquad \alpha_0 > 0 \,, \quad \alpha_1 > 0 \,, \quad 1 > \beta_1 \geq 0 \,. \tag{32}$$

The case $\alpha_1 = 0$ corresponds to a model with constant variance, which is of no interest in the current discussion.

An interesting property of the GARCH process is that its unconditional distribution is fat-tailed even with light-tailed (e. g., Gaussian) innovations, i. e., the distributional properties of the returns will not reflect those of the innovation (news) process. This has been known basically since [37,73], who showed that, even with normally distributed innovations, (G)ARCH processes do not have all their moments finite. For example, for the GARCH(1,1) model, [37] showed that, with $m \in \mathbb{N}$, the unconditional $(2m)$th moment of ϵ_t in (27) is finite if and only if

$$\mathbb{E}\left[(\alpha_1 \eta_t^2 + \beta_1)^m\right] < 1 \,, \tag{33}$$

which, as long as $\alpha_1 > 0$, will eventually be violated for all practically relevant distributions. The argument in [37] is based on the relation

$$\mathbb{E}\left[\sigma_t^{2m}\right] = \sum_{i=0}^{m} \binom{m}{i} \alpha_0^i \mathbb{E}\left[\left(\alpha_1 \eta_{t-1}^2 + \beta_1\right)^{m-i}\right] \mathbb{E}\left[\sigma_{t-1}^{2(m-i)}\right] \,, \tag{34}$$

which follows from (32). The coefficient of $\mathbb{E}\left[\sigma_{t-1}^{2m}\right]$ on the right-hand side of (34) is just the expression appearing in (33), and consequently the $(2m)$th unconditional moment cannot be finite if this exceeds unity. The heavy-tailedness of the GARCH process is sometimes also exemplified by means of its unconditional kurtosis measure (3), which is finite for the GARCH(1,1) model with Gaussian innovations iff $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$. Writing (34) down for $m = 2$, using (31) and substituting into (3) gives

$$\mathbb{K}\left[\epsilon_t\right] = \frac{3\left[1 - (\alpha_1 + \beta_1)^2\right]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3 \,,$$

as $\mathbb{E}\left[\epsilon_t^4\right] = 3\mathbb{E}\left[\sigma_t^4\right]$. [73] notes that "[m]any statistical procedures have been designed to be robust to large errors, but … none of this literature has made use of the fact that temporal clustering of outliers can be used to predict their occurrence and minimize their effects. This is exactly the approach taken by the ARCH model". Conditions for the existence of and expressions for higher-order moments of the GARCH(p, q) model can be found in [50,105,122,139]. The relation between the conditional and unconditional kurtosis of GARCH models was investigated in [15], see also [47] for related results.

A more precise characterization of the tails of GARCH processes has been developed by applying classical results about the tail behavior of solutions of stochastic difference equations as, for example, in [124]. We shall continue to concentrate on the GARCH(1,1) case, which admits relatively explicit results, and which has already been written as a first-order stochastic difference equation in (32). Iterating this,

$$\sigma_t^2 = \sigma_0^2 \prod_{i=1}^{t} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right) + \alpha_0 \left[1 + \sum_{k=1}^{t-1} \prod_{i=1}^{k} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right) \right] . \tag{35}$$

Nelson [171] has shown that the GARCH(1,1) process (32) has a strictly stationary solution, given by

$$\sigma_t^2 = \alpha_0 \left[1 + \sum_{k=1}^{\infty} \prod_{i=1}^{k} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right) \right] , \tag{36}$$

if and only if

$$\mathbb{E}\left[\log\left(\alpha_1 \eta_t^2 + \beta_1\right)\right] < 0 . \tag{37}$$

The keynote of the argument in [171] is the application of the strong law of large numbers to the terms of the form $\prod_{i=1}^{k}(\alpha_1\eta_{t-i}^2 + \beta_1) = \exp\{\sum_1^k \log\left(\alpha_1\eta_{t-i}^2 + \beta_1\right)\}$ in (35), revealing that (35) converges almost surely if (37) holds. Note that $\mathbb{E}\left[\log\left(\alpha_1\eta_t^2 + \beta_1\right)\right] < \log \mathbb{E}\left[\alpha_1\eta_t^2 + \beta_1\right] = \log\left(\alpha_1 + \beta_1\right)$, i. e., stationary GARCH processes need not be covariance stationary. Using (36) and standard moment inequalities, [171] further established that, in case of stationarity, $\mathbb{E}\left[|\epsilon_t|^p\right]$, $p > 0$, is finite if and only if $\mathbb{E}[(\alpha_1\eta_t^2 + \beta_1)^{p/2}] < 1$, which generalizes (33) to non-integer moments. It may now be supposed, and, building on the results of [90,124], has indeed been established by [167], that the tails of the marginal distribution of ϵ_t generated by a GARCH(1,1) process decay asymptotically in a Pareto-type fashion, i. e.,

$$P(|\epsilon_t| > x) \cong cx^{-\alpha} \quad \text{as } x \to \infty , \tag{38}$$

where the tail index α is the unique positive solution of the equation

$$h(\alpha) := \mathbb{E}\left[\left(\alpha_1\eta_t^2 + \beta_1\right)^{\alpha/2}\right] = 1 . \tag{39}$$

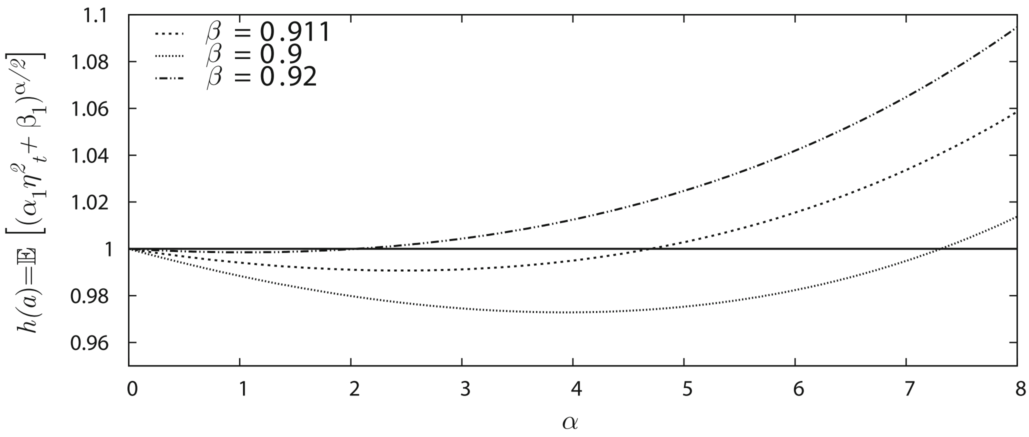
This follows from (28) along with the result that the tails of σ_t^2 and σ_t are asymptotically Paretian with tail indices $\alpha/2$ and α, respectively. For a discussion of technical conditions, see [167]. [167] also provides an expression for the constant c in (38), which is difficult to calculate explicitly, however. For the ARCH(1) model with Gaussian innovations, (39) becomes $(2\alpha_1)^{\alpha/2}\,\Gamma\left[(\alpha+1)/2\right]/\sqrt{\pi} = 1$, which has already been obtained by [63] and was foreshadowed in the work of [168]. The results reported above have been generalized in various directions, with qualitatively similar conclusions. The GARCH(p, q) case is treated in [29], while [140,141] consider various extensions of the standard GARCH(1,1) model.

Although the *unconditional* distribution of a GARCH model with Gaussian innovations has genuinely fat tails, it is often found in applications that the tails of empirical return distributions are even fatter than those implied by fitted Gaussian GARCH models, indicating that the *conditional* distribution, i. e., the distribution of η_t in (27), is likewise fat-tailed. Therefore, it has become standard practice to assume that the innovations η_t are also heavy tailed, although it has been questioned whether this is the best modeling strategy [199]. The most popular example of a heavy tailed innovation distribution is certainly the t considered in Subsect. "The Student t Distribution", which was introduced by [38] into the GARCH literature. Some authors have also found it beneficial to let the degrees of freedom parameter ν in (23) be time-varying, thus obtaining time-varying conditional fat-tailedness [45].

In the following, we shall briefly discuss a few GARCH(1,1) estimation results for the S&P500 series in order to compare the tails implied by these models with those from the semiparametric estimation procedures in Sect. "Empirical Evidence About the Tails". As distributions for the innovation process $\{\eta_t\}$, we shall consider the Gaussian, t, and the generalized error distribution (GED), which was introduced by [172] into the GARCH literature, see [128] for a recent contribution and asymmetric extensions. It has earlier been used in an unconditional context by [94] for the S&P500 returns. The density of the GED with mean zero and unit variance is given by

$$f_{\text{GED}}(x;\nu) = \frac{\lambda\nu}{2^{1/\nu+1}\,\Gamma(1/\nu)} \exp\left(-\frac{|\lambda x|^\nu}{2}\right) , \quad \nu > 0 , \tag{40}$$

where $\lambda = 2^{1/\nu}\sqrt{\Gamma(3/\nu)/\Gamma(1/\nu)}$. Parameter ν in (40) controls the thickness of the tails. For $\nu = 2$, we get the normal distribution, and a leptokurtic shape is obtained for $\nu < 2$. In the latter case, the tails of (40) are therefore thicker than those of the Gaussian, but they are not fat in the Pareto sense. However, even if one argues for Pareto tails of return distributions, use of (40) may be appropriate as a conditional distribution in GARCH models, because the power law already accompanies the volatility dynamics. To make the estimates of the parameter α_1 in (32) comparable, we also use the unit variance version of the t, which requires multiplying X in (23) by $\sqrt{(\nu-2)/\nu}$. Returns are modeled as $r_t = \mu + \epsilon_t$, where μ is a constant mean and ϵ_t is generated by (27) and (32). Parameter estimates, obtained by maximum-likelihood estimation, are provided in Table 2. In addition to the GARCH parameters in (32) and the shape parameters of the innovation distributions, Table 2 reports the log likelihood values and



Financial Economics, Fat-Tailed Distributions, Figure 8
The figure displays the function $h(\alpha)$, as defined in (39), for Gaussian η_t, $\alpha = 0.0799$ and various values of β_1. Note that $\hat{\alpha}_1 = 0.0799$ and $\hat{\beta}_1 = 0.911$ are the maximum likelihood estimates for the S&P500 returns, as reported in Table 2

the implied tail indices, $\hat{\alpha}$, which are obtained by solving (39) numerically. First note that all the GARCH models have considerably higher likelihood values than the iid models in Table 1, which highlights the importance of accounting for conditional heteroskedasticity. We can also conclude that the Gaussian assumption is still inadequate as a conditional distribution in GARCH models, as both the t and the GED achieve significantly higher likelihood values, and their estimated shape parameters indicate pronounced non-normalities. However, the degrees of freedom parameter of the t, ν, is somewhat increased in comparison to Table 1, as part of the leptokurtosis is now explained by the GARCH effects.

Compared to the nonparametric tail estimates obtained in Sect. "Empirical Evidence About the Tails", the tail index implied by the Gaussian GARCH(1,1) model turns out to be somewhat too high, while those of the more flexible models are both between 3 and 4 and therefore more in line with what has been found in Sect. "Empirical Evidence About the Tails". However, for all three models, the confidence intervals for α, as obtained from 1,000 simulations from the respective estimated GARCH processes, are rather wide, so that we cannot conclusively rule out the existence of the unconditional fourth (and even fifth) moment. The width of the confidence intervals reflects the fact that the implied tail indices are very sensitive to small variations in the underlying GARCH parameters. For example, if, in the GARCH model with conditional normality, we replace the estimate $\hat{\beta}_1 = 0.911$ with 0.9, the implied tail index is 7.31, and with $\beta_1 = 0.92$, we get $\alpha = 2.05$, which is close to an infinite variance. The situation is depicted in Fig. 8, showing $h(\alpha)$ in (39) for the different values of β_1. The shape of h follows generally from $h(0) = 1$, $h'(0) < 0$ by (37), $h'' > 0$, i. e., h is convex, and $\lim_{\alpha\to\infty} h(\alpha) = \infty$ as long as $P\left[(\alpha_1\eta_t^2 + \beta_1) > 1\right] > 0$, so that $h(\alpha) = 1$ has a unique positive solution. Note that both 0.9 and 0.92 are covered by $0.911 \pm 2 \times 0.009$, i. e., a 95% confidence interval for β_1. This shows that the GARCH-implied tail indices are rather noisy.

Alternatively, we may avoid precise assumptions about the distribution of the innovation process $\{\eta_t\}$ and rely on quasi maximum-likelihood results [138]. That is, we estimate the innovations by $\hat{\eta}_t = \hat{\epsilon}_t/\hat{\sigma}_t$, $t = 1, \ldots, 5,550$, where $\{\hat{\sigma}_t\}$ is the sequence of conditional standard deviations implied by the estimated Gaussian GARCH model, and then solve the sample analogue of (39), i. e., $T^{-1}\sum_{t=1}^{T}(\hat{\alpha}_1\hat{\eta}_t^2 + \hat{\beta}_1)^{\alpha/2} = 1$, a procedure theoretically justified in [32]. Doing so, we obtain $\hat{\alpha} = 2.97$, so that we recover the "universal cubic law". However, the 95% confidence interval, calculated from 1,000 GARCH simulations, where the innovation sequences are obtained by sampling with replacement from the $\hat{\eta}_t$-series, is (1.73, 4.80), which is still reconcilable with a finite fourth moment, and even with an infinite second moment. These results clearly underline the caveat brought out by [72] (p. 349), that "[t]here is no free lunch when it comes to [tail index] estimation".

Application to Value-at-Risk

In this section, we compare the models discussed in Sects. "Some Specific Distributions" and "Volatility Clustering and Fat Tails" on an economic basis by employing the Value-at-Risk (VaR) concept, which is a widely used measure to describe the downside risk of a financial position both in industry and in academia [118]. Consider a time

Financial Economics, Fat-Tailed Distributions, Table 2
GARCH parameter estimates

Distribution	$\hat{\mu}$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\nu}$	$\hat{\alpha}$	Loglik
Normal	0.059	0.012	0.080	0.911	—	4.70	−7271.7
	(0.011)	(0.002)	(0.008)	(0.009)		(3.20, 7.22)	
GED	0.063	0.007	0.058	0.936	1.291	3.95	−7088.2
	(0.010)	(0.002)	(0.007)	(0.008)	(0.031)	(2.52, 6.95)	
Symmetric *t*	0.063	0.006	0.051	0.943	6.224	3.79	−7068.1
	(0.010)	(0.002)	(0.006)	(0.007)	(0.507)	(2.38, 5.87)	

Shown are maximum-likelihood estimation results for GARCH(1,1) models, as given by (27) and (32), with different assumptions about the distribution of the innovations η_t in (27). Standard errors for the model parameters and 95% confidence intervals for the implied tail indices, $\hat{\alpha}$, are given in parentheses. "Loglik" is the value of the maximized log likelihood function.

series of portfolio returns, r_t, and an associated series of ex-ante VaR measures with target probability ξ, $\mathrm{VaR}_t(\xi)$. The $\mathrm{VaR}_t(\xi)$ implied by a model $\mathcal{M}$ is defined by

$$\Pr^{\mathcal{M}}_{t-1}\left(r_t < -\mathrm{VaR}_t(\xi)\right) = \xi \,, \tag{41}$$

where $\Pr^{\mathcal{M}}_{t-1}(\cdot)$ denotes a probability derived from model $\mathcal{M}$ using the information up to time $t-1$, and the negative sign in (41) is due to the convention of reporting VaR as a positive number. For an appropriately specified model, we expect $100 \times \xi\%$ of the observed return values not to exceed the (negative of the) respective VaR forecast. Thus, to assess the performance of the different models, we examine the percentage shortfall frequencies,

$$U_\xi = 100 \times \frac{x}{T} = 100 \times \hat{\xi} \,, \tag{42}$$

where T denotes the number of forecasts evaluated, x is the observed shortfall frequency, i. e., the number of days for which $r_t < -\mathrm{VaR}_t(\xi)$, and $\hat{\xi} = x/T$ is the empirical shortfall probability. If $\hat{\xi}$ is significantly less (higher) than ξ, then model $\mathcal{M}$ tends to overestimate (underestimate) the risk of the position. In the present application, in order to capture even the more extreme tail region, we focus on the target probabilities $\xi = 0.001, 0.0025, 0.005, 0.01, 0.025$, and 0.05.

To formally test whether a model correctly estimates the risk (according to VaR) inherent in a given financial position, that is, whether the empirical shortfall probability, $\hat{\xi}$, is statistically indistinguishable from the nominal shortfall probability, ξ, we use the likelihood ratio test [133]

$$\mathrm{LRT}_{\mathrm{VaR}} = -2\left\{x \log\frac{\xi}{\hat{\xi}} + (T-x)\log\frac{1-\xi}{1-\hat{\xi}}\right\} \overset{\mathrm{asy}}{\sim} \chi^2(1) \,. \tag{43}$$

On the basis of the first 1,000 return observations, we calculate one-day-ahead VaR measures based on parameter estimates obtained from an expanding data window, where the parameters are updated every day. Thus we get, for each model, 4,550 one-day-ahead out-of-sample VaR measures.

Table 3 reports the realized one-day-ahead percentage shortfall frequencies for the different target probabilities, ξ, as given above. The upper panel of the table shows the results for the unconditional distributions discussed in Sect. "Some Specific Distributions". The results clearly show that the normal distribution strongly *underestimates* ($\hat{\xi} > \xi$) the downside risk for the lower target probabilities, while the Cauchy as well as the alpha stable distributions tend to significantly *overestimate* ($\hat{\xi} < \xi$) the tails. This is in line with what we have observed from the empirical density plots presented in Fig. 7, which, in contrast to the out-of-sample VaR calculations, are based on estimates for the entire sample. Interestingly, the finite normal mixture distribution also tends to overestimate the risk at the lower VaR levels, leading to a rejection of correct coverage for almost all target probabilities. In contrast, the HYP distribution, whose empirical tails have been very close to those of the normal mixture in-sample (see Fig. 7), nicely reproduces the target probabilities, as does the VG distribution.

Similarly to the log likelihood results presented in Subsect. "Empirical Comparison" the Value-at-Risk evaluation does not allow for a clear distinction between the different t distributions, the GH and the NIG distribution. Similar to the Cauchy and the stable, they all tend to overestimate the more extreme target probabilities, while they imply too large shortfall probabilities at the five percent quantile.

The fact that most unconditional distributional models tend to overestimate the risk at the lower target probabil-

Financial Economics, Fat-Tailed Distributions, Table 3
Backtesting Value-at-Risk measures

Unconditional Distributional Models						
Distribution	$U_{0.001}$	$U_{0.0025}$	$U_{0.005}$	$U_{0.01}$	$U_{0.025}$	$U_{0.05}$
GH	0.04	0.11**	0.24***	0.73*	2.70	5.89***
t_{GH}	0.07	0.11**	0.22***	0.75*	2.75	5.96***
t_{JF}	0.04	0.11**	0.31**	0.88	2.64	5.32
t_{AC}	0.04	0.11**	0.26**	0.84	2.48	5.16
t_{FS}	0.07	0.13*	0.33**	0.95	2.77	5.38
Symmetric t	0.07	0.15	0.31**	0.92	3.08**	6.35***
NIG	0.07	0.15	0.26**	0.70**	2.35	5.34
HYP	0.13	0.24	0.51	0.95	2.50	5.16
VG	0.13	0.24	0.51	0.92	2.46	5.10
Alpha stable	0.04	0.11**	0.33**	0.75*	2.44	4.90
Finite mixture ($k = 2$)	0.04	0.07***	0.11***	0.37***	2.99**	6.40***
Cauchy	0.00***	0.00***	0.00***	0.00***	0.09***	0.88***
Normal	0.48***	0.64***	0.97***	1.36**	2.44	4.02***

GARCH(1,1) Models						
Distribution	$U_{0.001}$	$U_{0.0025}$	$U_{0.005}$	$U_{0.01}$	$U_{0.025}$	$U_{0.05}$
Normal	0.40***	0.66***	0.92***	1.36**	2.95*	4.57
GED	0.20*	0.33	0.44	0.79	2.48	4.79
Symmetric t	0.11	0.26	0.40	0.92	2.86	5.45

The table shows the realized one-day-ahead percentage shortfall frequencies, U_ξ, for given target probabilities, ξ, as defined in (42). Asterisks *, ** and *** indicate significance at the 10%, 5% and 1% levels, respectively, as obtained from the likelihood ratio test (43).

ities may be due to our use of an expanding data window and the impact of the "Black Monday", where the index decreased by more than 20%, at the beginning of our sample period. In this regard, the advantages of accounting for time-varying volatility via a GARCH(1,1) structure may become apparent, as this model allows the more recent observations to have much more impact on the conditional density forecasts.

In fact, by inspection of the results for the GARCH models, as reported in the lower part of Table 3, it turns out that the GARCH(1,1) model with a normal distribution strongly underestimates the empirical shortfall probabilities at all levels except the largest (5%). However, considering a GED or t distribution for the return innovations within the GARCH model provides accurate estimates of downside risks.

To further discriminate between the GARCH processes and the iid models, tests for *conditional* coverage may be useful, which are discussed in the voluminous VaR literature (e. g., [53]).

Finally, we point out that the current application is necessarily of an illustrative nature. In particular, if the data generating process is not constant but evolves slowly over time and/or is subject to abrupt structural breaks, use of a rolling data window will be preferred to an expanding window.

Future Directions

As highlighted in the previous sections, there exists a plethora of different and well-established approaches for modeling the tails of univariate financial time series. However, on the multivariate level the number of models and distributions is still very limited, although the joint modeling of multiple asset returns is crucial for portfolio risk management and allocation decisions. The problem is then to model the dependencies between financial assets. In the literature, this problem has been tackled, for example, by means of multivariate extensions of the mean-variance mixture (18) [19], multivariate GARCH models [30], regime-switching models [11], and copulas [51]. The problem is particularly intricate if the number of assets to be considered is large, and much work remains to understand and properly model their dependence structure.

It is also worth mentioning that the class of GARCH processes, due to its interesting conditional and unconditional distributional properties, has been adopted, for example, in the signal processing literature [3,52,55], and it is to be expected that it will be applied in other fields in the future.

Bibliography

Primary Literature

1. Aas K, Haff IH (2006) The generalized hyperbolic skew Student's *t*-distribution. J Financial Econom 4:275–309
2. Abhyankar A, Copeland LS, Wong W (1995) Moment condition failure in high frequency financial data: Evidence from the S&P 500. Appl Econom Lett 2:288–290
3. Abramson A, Cohen I (2006) State smoothing in Markov-switching time-frequency GARCH models. IEEE Signal Process Lett 13:377–380
4. Akgiray V, Booth GG (1988) The stable-law model of stock returns. J Bus Econ Stat 6:51–57
5. Alexander C, Lazar E (2006) Normal mixture GARCH(1,1). Applications to exchange rate modelling. J Appl Econom 21:307–336
6. Alexander SS (1961) Price movements in speculative markets: Trends or random walks. Ind Manag Rev 2:7–25
7. Andersen TG, Bollerslev T (1998) Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. Int Econ Rev 39:885–905
8. Andersen TG, Bollerslev T, Diebold FX, Ebens H (2001) The distribution of realized stock return volatility. J Financial Econ 61:43–76
9. Andersen TG, Bollerslev T, Diebold FX, Labys P (2001) The distribution of realized exchange rate volatility. J Am Stat Assoc 96:42–55
10. Andersson J (2001) On the normal inverse Gaussian stochastic volatility model. J Bus Econ Stat 19:44–54
11. Ang A, Chen J (2002) Asymmetric correlations of equity portfolios. J Financial Econ 63:443–494
12. Atkinson AC (1982) The simulation of generalized inverse Gaussian and hyperbolic random variables. SIAM J Sci Stat Comput 3:502–515
13. Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. J R Stat Soc Ser B 65:367–389
14. Bachelier L (1964) Theory of speculation. In: Cootner PH (ed) The random character of stock market prices. MIT Press, Cambridge, pp 17–75
15. Bai X, Russell JR, Tiao GC (2003) Kurtosis of GARCH and stochastic volatility models with non-normal innovations. J Econom 114:349–360
16. Balanda KP, MacGillivray HL (1988) Kurtosis: A critical review. Am Stat 42:111–119
17. Barndorff-Nielsen OE (1977) Exponentially decreasing distributions for the logarithm of particle size. Proc R Soc Lond Ser A 353:401–419
18. Barndorff-Nielsen OE (1988) Processes of normal inverse Gaussian type. Finance Stoch 2:41–68
19. Barndorff-Nielsen OE (1997) Normal inverse Gaussian distributions and stochastic volatility modelling. Scand J Stat 24:1–13
20. Barndorff-Nielsen OE, Halgreen C (1977) Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Probab Theory Relat Fields 38:309–311
21. Barndorff-Nielsen OE, Prause K (2001) Apparent scaling. Finance Stoch 5:103–113
22. Barndorff-Nielsen OE, Shephard N (2001) Normal modified stable processes. Theory Prob Math Stat 65:1–19
23. Barndorff-Nielsen OE, Shephard N (2002) Estimating quadratic variation using realized variance. J Appl Econom 17:457–477
24. Barndorff-Nielsen OE, Shephard N (2007) Financial volatility in continuous time. Cambridge University Press
25. Barndorff-Nielsen OE, Stelzer R (2005) Absolute moments of generalized hyperbolic distributions and approximate scaling of normal inverse Gaussian Lévy processes. Scand J Stat 32:617–637
26. Barndorff-Nielsen OE, Kent J, Sørensen M (1982) Normal variance-mean mixtures and z distributions. Int Stat Rev 50:145–159
27. Barndorff-Nielsen OE, Blæsild P, Jensen JL, Sørensen M (1985) The fascination of sand. In: Atkinson AC, Fienberg SE (eds) A celebration of statistics. Springer, Berlin, pp 57–87
28. Barnea A, Downes DH (1973) A reexamination of the empirical distribution of stock price changes. J Am Stat Assoc 68:348–350
29. Basrak B, Davis RA, Mikosch T (2002) Regular variation of GARCH processes. Stoch Process Appl 99:95–115
30. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109
31. Beale EML, Mallows CL (1959) Scale mixing of symmetric distributions with zero mean. Ann Math Stat 30:1145–1151
32. Berkes I, Horváth L, Kokoszka P (2003) Estimation of the maximal moment exponent of a GARCH(1,1) sequence. Econom Theory 19:565–586
33. Bibby BM, Sørensen M (1997) A hyperbolic diffusion model for stock prices. Finance Stoch 1:25–41
34. Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Cambridge University Press
35. Blanchard OJ, Watson MW (1982) Bubbles, rational expectations, and financial markets. In: Wachtel P (ed) Crises in the economic and financial structure. Lexington Books, Lexington, pp 295–315
36. Blattberg RC, Gonedes NJ (1974) A comparison of the stable and Student distributions as statistical models for stock prices. J Bus 47:244–280
37. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31:307–327
38. Bollerslev T (1987) A conditionally heteroskedastic time series model for speculative prices and rates of return. Rev Econ Stat 69:542–547
39. Bollerslev T, Kretschmer U, Pigorsch C, Tauchen G (2007) A discrete-time model for daily S&P500 returns and realized variations: Jumps and leverage effects. J Econom (in press)
40. Boothe P, Glassman D (1987) The statistical distribution of exchange rates. J Int Econ 22:297–319
41. Bouchaud JP, Potters M (2000) Theory of financial risks. From statistical physics to risk management. Cambridge University Press, Cambridge

42. Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann Math Stat 29:610–611
43. Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. J Multivar Anal 79:99–113
44. Broca DS (2004) Mixture distribution models of Indian stock returns: An empirical comparison. Indian J Econ 84:525–535
45. Brooks C, Burke SP, Heravi S, Persand G (2005) Autoregressive conditional kurtosis. J Financial Econom 3:399–421
46. Campbell J, Lo AW, MacKinlay AC (1997) The econometrics of financial markets. Princeton University Press, Princeton
47. Carnero MA, Peña D, Ruiz E (2004) Persistence and kurtosis in GARCH and stochastic volatility models. J Financial Econom 2:319–342
48. Carr P, Geman H, Madan DB, Yor M (2002) The fine structure of asset returns: An empirical investigation. J Bus 75:305–332
49. Chambers JM, Mallows CL, Stuck BW (1976) A method for simulating stable random variables. J Am Stat Assoc 71:340–344
50. Chen M, An HZ (1998) A note on the stationarity and the existence of moments of the GARCH model. Stat Sin 8:505–510
51. Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. Wiley, New York
52. Cheung YM, Xu L (2003) Dual multivariate auto-regressive modeling in state space for temporal signal separation. IEEE Trans Syst Man Cybern B: 33:386–398
53. Christoffersen PF, Pelletier D (2004) Backtesting value-at-risk: A duration-based approach. J Financial Econom 2:84–108
54. Clark PK (1973) A subordinated stochastic process model with finite variance for speculative prices. Econometrica 41:135–155
55. Cohen I (2004) Modeling speech signals in the time-frequency domain using GARCH. Signal Process 84:2453–2459
56. Cont R, Tankov P (2004) Financial modelling with jump processes. Chapman & Hall, Boca Raton
57. Corsi F, Mittnik S, Pigorsch C, Pigorsch U (2008) The volatility of realized volatility. Econom Rev 27:46–78
58. Cotter J (2005) Tail behaviour of the euro. Appl Econ 37:827–840
59. Dacorogna MM, Müller UA, Pictet OV, de Vries CG (2001) Extremal forex returns in extremely large data sets. Extremes 4:105–127
60. Dagpunar JS (1989) An easily implemented generalised inverse Gaussian generator. Commun Stat Simul 18:703–710
61. Danielsson J, de Vries CG (1997) Tail index and quantile estimation with very high frequency data. J Empir Finance 4:241–257
62. Danielsson J, de Haan L, Peng L, de Vries CG (2001) Using a bootstrap method to choose the sample fraction in tail index estimation. J Multivar Anal 76:226–248
63. de Haan L, Resnick SI, Rootzén H, de Vries CG (1989) Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. Stoch Process Appl 32:213–234
64. de Vries CG (1994) Stylized facts of nominal exchange rate returns. In: van der Ploeg F (ed) The handbook of international macroeconomics. Blackwell, Oxford, pp 348–389
65. Doganoglu T, Mittnik S (1998) An approximation procedure for asymmetric stable Paretian densities. Comput Stat 13:463–475
66. DuMouchel WH (1973) On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. Ann Stat 1:948–957
67. DuMouchel WH (1983) Estimating the stable index α in order to measure tail thickness: A critique. Ann Stat 11:1019–1031
68. Dyson FJ (1943) A note on kurtosis. J R Stat Soc 106:360–361
69. Eberlein E, von Hammerstein EA (2004) Generalized hyperbolic and inverse Gaussian distributions: Limiting cases and approximation of processes. In: Dalang RC, Dozzi M, Russo F (eds) Seminar on stochastic analysis, random fields and applications IV. Birkhäuser, Basel, pp 221–264
70. Eberlein E, Keller U (1995) Hyperbolic distributions in finance. Bernoulli 1:281–299
71. Eberlein E, Keller U, Prause K (1998) New insights into smile, mispricing, and value at risk: The hyperbolic model. J Bus 71:371–405
72. Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
73. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1006
74. Fama EF (1963) Mandelbrot and the stable Paretian hypothesis. J Bus 36:420–429
75. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–105
76. Fama EF (1976) Foundations of finance. Basic Books, New York
77. Fama EF, Roll R (1971) Parameter estimates for symmetric stable distributions. J Am Stat Assoc 66:331–338
78. Farmer JD, Lillo F (2004) On the origin of power-law tails in price fluctuations. Quantit Finance 4:C7–C11
79. Feller W (1950) An introduction to probability theory and its applications I. Wiley, New York
80. Feller W (1971) An introduction to probability theory and its applications II. Wiley, New York
81. Fernández C, Steel MFJ (1998) On Bayesian modeling of fat tails and skewness. J Am Stat Assoc 93:359–371
82. Finucan HM (1964) A note on kurtosis. J R Stat Soc Ser B 26:111–112
83. Gabaix X, Gopikrishnan P, Plerou V, Stanley E (2007) A unified econophysics explanation for the power-law exponents of stock market activity. Physica A 382:81–88
84. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423:267–270
85. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2006) Institutional investors and stock market volatility. Quart J Econ 121:461–504
86. Galbraith JW, Zernov S (2004) Circuit breakers and the tail index of equity returns. J Financial Econom 2:109–129
87. Gallant AR, Tauchen G (1996) Which moments to match? Econom Theory 12:657–681
88. Gerber HU, Shiu ESW (1994) Option pricing by Esscher transforms. Trans Soc Actuar 46:99–140
89. Ghose D, Kroner KF (1995) The relationship between GARCH and symmetric stable processes: Finding the source of fat tails in financial data. J Empir Finance 2:225–251
90. Goldie CM (1991) Implicit renewal theory and tails of solutions of random equations. Ann Appl Probab 1:126–166
91. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse cubic law for the distribution of stock price variations. Eur Phys J B 3:139–140

92. Gopikrishnan P, Plerou V, Amaral LAN, Meyer M, Stanley HE (1999) Scaling of the distribution of fluctuations of financial market indices. Phys Rev E 60:5305–5316
93. Gourieroux C, Jasiak J (1998) Truncated maximum likelihood, goodness of fit tests and tail analysis. Working Paper, CREST
94. Gray JB, French DW (1990) Empirical comparisons of distributional models for stock index returns. J Bus Finance Account 17:451–459
95. Gut A (2005) Probability: A graduate course. Springer, New York
96. Haas M, Mittnik S, Paolella MS (2004) Mixed normal conditional heteroskedasticity. J Financial Econom 2:211–250
97. Haas M, Mittnik S, Paolella MS (2004) A new approach to Markov-switching GARCH models. J Financial Econom 2: 493–530
98. Hagerman RL (1978) More evidence on the distribution of security returns. J Finance 33:1213–1221
99. Hall JA, Brorsen W, Irwin SH (1989) The distribution of futures prices: A test of the stable Paretian and mixture of normals hypotheses. J Financial Quantit Anal 24:105–116
100. Hall P (1982) On some simple estimates of an exponent of regular variation. J R Stat Soc Ser B 44:37–42
101. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
102. Hansen BE (1994) Autoregressive conditional density estimation. Int Econ Rev 35:705–730
103. Harrison P (1996) Similarities in the distribution of stock market price changes between the eighteenth and twentieth centuries. J Bus 71:55–79
104. Hazan A, Landsman Z, Makov UE (2003) Robustness via a mixture of exponential power distributions. Comput Stat Data Anal 42:111–121
105. He C, Teräsvirta T (1999) Fourth moment structure of the GARCH(p, q) process. Econom Theory 15:824–846
106. Hill BM (1975) A simple general approach to inference about the tail of a distribution. Ann Stat 3:1163–1174
107. Hols MCAB, de Vries CG (1991) The limiting distribution of extremal exchange rate returns. J Appl Econom 6:287–302
108. Hsieh DA (1989) Modeling heteroskedasticity in daily foreign-exchange rates. J Bus Econ Stat 7:307–317
109. Hsu DA, Miller RB, Wichern DW (1974) On the stable Paretian behavior of stock-market prices. J Am Stat Assoc 69:108–113
110. Huisman R, Koedijk KG, Kool CJM, Palm F (2001) Tail-index estimates in small samples. J Bus Econ Stat 19:208–216
111. Huisman R, Koedijk KG, Kool CJM, Palm F (2002) The tail-fatness of FX returns reconsidered. Economist 150:299–312
112. Hyung N, de Vries CG (2005) Portfolio diversification effects of downside risk. J Financial Econom 3:107–125
113. Jansen DW, de Vries CG (1991) On the frequency of large stock returns: Putting booms and busts into perspective. Rev Econ Stat 73:18–24
114. Jensen MB, Lunde A (2001) The NIG-S&ARCH model: A fat-tailed, stochastic, and autoregressive conditional heteroskedastic volatility model. Econom J 4:319–342
115. Jondeau E, Rockinger M (2003) Testing for differences in the tails of stock-market returns. J Empir Finance 10:559–581
116. Jones MC, Faddy MJ (2003) A skew extension of the t-distribution, with applications. J R Stat Soc Ser B 65:159–174
117. Jørgensen B (1982) Statistical properties of the generalized inverse gaussian distribution. Springer, Berlin
118. Jorion P (2006) Value at risk: The new benchmark for controlling derivatives risk. McGraw-Hill, New York
119. Kaizoji T, Kaizoji M (2003) Empirical laws of a stock price index and a stochastic model. Adv Complex Syst 6:303–312
120. Kanter M (1975) Stable densities under change of scale and total variation inequalities. J Am Stat Assoc 3:697–707
121. Kaplansky I (1945) A common error concerning kurtosis. J Am Stat Assoc 40:259
122. Karanasos M (1999) The second moment and the autocovariance function of the squared errors of the GARCH model. J Econom 90:63–76
123. Kearns P, Pagan A (1997) Estimating the density tail index for financial time series. Rev Econ Stat 79:171–175
124. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Math 131:207–248
125. Koedijk KG, Schafgans MMA, de Vries CG (1990) The tail index of exchange rate returns. J Int Econ 29:93–108
126. Koedijk KG, Stork PA, de Vries CG (1992) Differences between foreign exchange rate regimes: The view from the tails. J Int Money Finance 11:462–473
127. Kogon SM, Williams DB (1998) Characteristic function based estimation of stable distribution parameters. In: Adler RJ, Feldman RE, Taqqu MS (eds) A practical guide to heavy tails: Statistical techniques and applications. Birkhäuser, Basel, pp 311–335
128. Komunjer I (2007) Asymmetric power distribution: Theory and applications to risk management. J Appl Econom 22:891–921
129. Kon SJ (1984) Models of stock returns: A comparison. J Finance 39:147–165
130. Koponen I (1995) Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. Phys Rev E 52:1197–1199
131. Koutrouvelis IA (1980) Regression-type estimation of the parameters of stable laws. J Am Stat Assoc 75:918–928
132. Küchler U, Neumann K, Sørensen M, Streller A (1999) Stock returns and hyperbolic distributions. Math Comput Model 29:1–15
133. Kupiec PH (1995) Techniques for verifying the accuracy of risk management models. J Deriv 3:73–84
134. Laherrère J, Sornette D (1998) Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. Eur Phys J B 2:525–539
135. Lau AHL, Lau HS, Wingender JR (1990) The distribution of stock returns: New evidence against the stable model. J Bus Econ Stat 8:217–223
136. Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York
137. LeBaron B (2001) Stochastic volatility as a simple generator of apparent financial power laws and long memory. Quantit Finance 1:621–631
138. Lee SW, Hansen BE (1994) Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. Econom Theory 10:29–52
139. Ling S, McAleer M (2002) Necessary and sufficient moment conditions for the GARCH(r, s) and asymmetric power GARCH(r, s) models. Econom Theory 18:722–729

140. Liu JC (2006) On the tail behaviors of a family of GARCH processes. Econom Theory 22:852–862
141. Liu JC (2006) On the tail behaviors of Box–Cox transformed threshold GARCH(1,1) processes. Stat Probab Lett 76:1323–1330
142. Longin FM (1996) The asymptotic distribution of extreme stock market returns. J Bus 69:383–408
143. Longin FM (2005) The choice of the distribution of asset returns: How extreme value theory can help? J Bank Finance 29:1017–1035
144. Loretan M, Phillips PCB (1994) Testing the covariance stationarity of heavy-tailed time series. J Empir Finance 1:211–248
145. Lux T (1996) The stable Paretian hypothesis and the frequency of large returns: An examination of major German stocks. Appl Financial Econ 6:463–475
146. Lux T (2000) On moment condition failure in German stock returns: An application of recent advances in extreme value statistics. Empir Econ 25:641–652
147. Lux T (2001) The limiting extremal behaviour of speculative returns: An analysis of intra-daily data from the Frankfurt stock exchange. Appl Financial Econ 11:299–315
148. Lux T, Sornette D (2002) On rational bubbles and fat tails. J Money Credit Bank 34:589–610
149. Madan DB, Carr PP, Chang EC (1998) The variance gamma process and option pricing. Eur Finance Rev 2:79–105
150. Madan DB, Milne F (1991) Option pricing with v. g. martingale components. Math Finance 1:39–55
151. Madan DB, Seneta E (1990) The variance gamma (v. g.) model for share market returns. J Bus 63(4):511–524
152. Malevergne Y, Pisarenko V, Sornette D (2005) Empirical distributions of stock returns: Between the stretched exponential and the power law? Quantit Finance 5:379–401
153. Malevergne Y, Pisarenko V, Sornette D (2006) On the power of generalized extreme value (GEV) and generalized Pareto distribution (GDP) estimators for empirical distributions of stock returns. Appl Financial Econ 16:271–289
154. Mandelbrot B (1963) New methods in statistical economics. J Polit Econ 71:421–440
155. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36:394–419
156. Mandelbrot B (1967) The variation of some other speculative prices. J Bus 40:393–413
157. Mantegna RN, Stanley HE (1994) tic process with ultraslow convergence to a Gaussian: The truncated Lévy flight. Phys Rev Lett 73:2946–2949
158. Marsaglia G, Marshall AW, Proschan F (1965) Moment crossings as related to density crossings. J R Stat Soc Ser B 27:91–93
159. Mason DM (1982) Laws of large numbers for sums of extreme values. Ann Probab 10:754–764
160. Matia K, Amaral LAN, Goodwin SP, Stanley HE (2002) Different scaling behaviors of commodity spot and future prices. Phys Rev E 66:045103
161. Matia K, Pal M, Salunkay H, Stanley HE (2004) Scale-dependent price fluctuations for the Indian stock market. Europhys Lett 66:909–914
162. McCulloch JH (1997) Measuring tail thickness to estimate the stable index α: A critique. J Bus Econ Stat 15:74–81
163. McCulloch JH (1986) Simple consistent estimators of stable distribution parameters. Commun Stat Simul 15:1109–1136
164. McDonald JB (1996) Probability distributions for financial models. In: Maddala GS, Rao CR (eds) Handbook of statistics 14: Statistical methods in finance. Elsevier, Amsterdam, pp 427–461
165. McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
166. Mikosch T, Stărică C (2004) Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Rev Econ Stat 86:378–390
167. Mikosch T, Stărică C (2000) Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. Ann Stat 28:1427–1451
168. Milhøj A (1985) The moment structure of ARCH processes. Scand J Stat 12:281–292
169. Mittnik S, Rachev ST (1993) Modeling asset retunrs with alternative stable distributions. Econom Rev 12:261–330
170. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Analysis of high-resolution foreign exchange data of USD-JPY for 13 years. Physica A 324:296–302
171. Nelson DB (1990) Stationarity and persistence in the GARCH(1,1) model. Econom Theory 6:318–334
172. Nelson DB (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
173. Nelson DB, Cao CQ (1992) Inequality constraints in the univariate GARCH model. J Bus Econ Stat 10:229–235
174. Newcomb S (1980) A generalized theory of the combination of observations so as to obtain the best result. In: Stigler SM (ed) American contributions to mathematical statistics in the nineteenth century, vol. 2. Arno, New York, pp. 343–366
175. Nolan JP (1997) Numerical calculation of stable densities and distribution functions. Commun Stat Stoch Models 13:759–774
176. Officer RR (1972) The distribution of stock returns. J Am Stat Assoc 67:807–812
177. Omran MF (1997) Moment condition failure in stock returns: UK evidence. Appl Math Finance 4:201–206
178. Osborne MFM (1959) Brownian motion in the stock market. Oper Res 7:145–173
179. Peiró A (1994) The distribution of stock returns: International evidence. Appl Financial Econ 4:431–439
180. Perry PR (1983) More evidence on the nature of the distribution of security returns. J Financial Quantit Anal 18:211–221
181. Plerou V, Gopikrishnan P, Amaral LAN, Meyer M, Stanley HE (1998) Scaling of the distribution of price fluctuations of individual companies. Phys Rev E 60:6519–6529
182. Plerou V, Gopikrishnan P, Gabaix X, Stanley E (2004) On the origin of power-law fluctuations in stock prices. Quantit Finance 4:C11–C15
183. Pólya G, Szegö G (1976) Problems and theorems in analysis II. Springer, Berlin
184. Praetz PD (1972) The distribution of share price changes. J Bus 45:49–55
185. Prause K (1999) The generalized hyperbolic model: Estimation, financial derivatives, and risk measures. Ph D thesis, Albert-Ludwigs-Universität Freiburg i. Br.
186. Press SJ (1972) Estimation in univariate and multivariate stable distributions. J Am Stat Assoc 67:842–846
187. Raible S (2000) Lévy processes in finance: Theory, numerics, and empirical facts. Ph D thesis, Albert-Ludwigs-Universität Freiburg i. Br.

188. Resnick SI (1987) Extreme values, regular variation, and point processes. Springer, New York
189. Resnick SI (1997) Heavy tail modeling and teletraffic data. Ann Stat 25:1805–1849
190. Resnick SI, Stărică C (1998) Tail index estimation for dependent data. Ann Appl Probab 8:1156–1183
191. Rydberg TH (1999) Generalized hyperbolic diffusion processes with applications in finance. Math Finance 9:183–201
192. Samorodnitsky G, Taqqu MS (1994) Stable non-gaussian random processes: Stochastic models with infinite variance. Chapman & Hall, New York
193. Sato KI (1999) Lévy processes and infinitely divisible distributions. Cambridge University Press, Cambridge
194. Schoutens W (2003) Lévy processes in finance: Pricing financial derivatives. Wiley, New York
195. Seneta E (1976) Regularly varying functions. Springer, Berlin
196. Sigman K (1999) Appendix: A primer on heavy-tailed distributions. Queueing Syst 33:261–275
197. Silva AC, Prange RE, Yakovenko VM (2004) Exponential distribution of financial returns at mesoscopic time lags: A new stylized fact. Physica A 344:227–235
198. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, New York
199. Tsiakas I (2006) Periodic stochastic volatility and fat tails. J Financial Econom 4:90–135
200. Turner CM, Startz R, Nelson CR (1989) A Markov model of heteroskedasticity, risk, and learning in the stock market. J Financial Econ 25:3–22
201. Werner T, Upper C (2004) Time variation in the tail behavior of bund future returns. J Future Markets 24:387–398
202. Weron R (2001) Levy-stable distributions revisited: Tail index $>$ 2 does not exclude the Levy-stable regime. Int J Mod Phys C 12:209–223

Books and Reviews

Bollerslev T, Engle RF, Nelson DB (1994) ARCH models. In: Engle RF, McFadden DL (eds) Handbook of econometrics IV. Elsevier, Amsterdam, pp 2959–3038

Borak S, Härdle W, Weron R (2005) Stable distributions. In: Cizek P, Härdle W, Weron R (eds) Statistical tools for finance and insurance. Springer, Berlin, pp 21–44

Janicki A, Weron A (1993) Simulation and chaotic behavior of α-stable stochastic processes. Dekker, New York

McCulloch JH (1996) Financial applications of stable distributions. In: Maddala GS, Rao CR (eds) Handbook of statistics 14: Statistical methods in finance. Elsevier, Amsterdam, pp 393–425

Mikosch T (2004) Modeling dependence and tails of financial time series. In: Finkenstädt B, Rootzén H (eds) Extreme values in finance, telecommunications, and the environment. Chapman & Hall, Boca Raton, pp 185–286

Mittnik S, Rachev ST, Paolella MS (1998) Stable paretian modeling in finance: Some empirical and theoretical aspects. In: Adler RJ, Feldman RE, Taqqu MS (eds) A practical guide to heavy tails: Statistical techniques and applications. Birkhäuser, Basel, pp 79–110

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Pagan A (1996) The econometrics of financial markets. J Empir Finance 3:15–102

Palm FC (1996) GARCH models of volatility. In: Maddala GS, Rao CR (eds) Handbook of statistics 14. Elsevier, Amsterdam, pp 209–240

Rachev ST, Mittnik S (2000) Stable paretian models in finance. Wiley, Chichester

Zolotarev VM (1986) One-dimensional stable distributions. AMS, Providence

Financial Economics, Non-linear Time Series in

Terence C. Mills[1], Raphael N. Markellos[1,2]
[1] Department of Economics, Loughborough University, Loughborough, UK
[2] Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece

Article Outline

Glossary

Arbitrage The possibility of producing a riskless profit by exploiting price differences between identical or linked assets.

Market efficiency A market is called efficient when all available information is reflected accurately, instantly and fully in the prices of traded assets. Depending on the definition of the available information set, private, public or that contained in historical prices, market efficiency is considered as strong, semi-strong or weak, respectively. The market price of an asset in an efficient market is an unbiased estimate of its true value. Systematic excess profits, which cannot be justified on the basis of the underlying risk, are not possible in such a market.

Martingale The term was originally used to describe a particular gambling strategy in which the stake is doubled following a losing bet. In probability theory it refers to a stochastic process that is a mathematical model of 'fair play'. This has been one of the most widely assumed processes for financial prices. It implies that the best forecast for tomorrow's price is simply today's price or, in other words, that the expected difference between any two successive prices is zero. Assuming a positive (negative) expected difference leads to the more general and realistic class of submartingale (supermartingale) processes. The martingale process implies that price differences are serially uncorrelated and that univariate linear time series models of prices have no forecasting value. However, martingales do not preclude the potential usefulness of nonlinear models in predicting the evolution of higher moments, such as the variance. The efficient market hypothesis is often incorrectly equated to the so-called random walk hypothesis, which roughly states that financial prices are martingales.

Option A call (put) option is a contractual agreement which gives the holder the right to buy (sell) a specified quantity of the underlying asset, within a specified period of time, at a price that is agreed when the contract is executed. Options are derivative assets since their value is based upon the variation in the underlying, which is typically the price of some asset such as a stock, commodity, bond, etc. Other basic types of derivatives include futures, forwards and swaps. An option is real, in contrast to financial, when the corresponding right refers to some business decision, such as the right to build a factory.

Portfolio theory The study of how resources should be optimally allocated between alternative investments on the basis of a given time investment horizon and a set of preferences.

Systematic risk Reflects the factors affecting all securities or firms in an economy. It cannot be reduced by diversification and it is also known as market risk. In the context of one of the most popular financial models, the Capital Asset Pricing Model (CAPM), systematic risk is measured by the beta coefficient.

Unsystematic risk This is the part of risk that is unique to a particular security or firm and can be reduced through diversification. This risk cannot be explained on the basis of fluctuations in the market as whole and it is also known as residual or idiosyncratic risk.

Volatility A measure of overall risk for an asset or portfolio which represents the sum of systematic and unsystematic risk. While several different approaches have been proposed for approximating this unobservable variable, the simplest one is based on the annualized standard deviation estimated using a historical sample of daily returns.

Definition of the Subject

Financial economics is the branch of economic science that deals with how groups of agents, such as households, firms, investors, creditors and economies as a whole, allocate and exchange financial resources in the context of markets. A wide variety of problems and applications fall within this broad subject area, including asset pricing, portfolio optimization, market efficiency, capital budgeting, interest and exchange rate modeling, risk management, forecasting and trading, market microstructure and behavioral finance. It is a highly quantitative and empirical discipline which draws its theoretical foundations and tools primarily from economics, mathematics and econometrics. Academic research in this area has flourished over the past century in line with the growing importance of financial markets and assets for the everyday life of corporations and individuals (for a historical overview of financial economics, see [69]). Consequently, at least 6 out of the 39 Nobel prizes in Economics have been awarded for research undertaken in areas related to financial economics. The close relationship between finance and time series analysis became widely apparent when Sir Clive W.J. Granger and Robert F. Engle III jointly received the 2003 Nobel Prize. Their work in time series econometrics has had a profound impact both on academic research and on the practice of finance. In particular, the ARCH model, first proposed by Engle [27] for modeling the variability of inflation, is today one of the most well known and important applications of a nonlinear time series model in finance. We should also acknowledge the Nobel Prize received by Robert C. Merton and Myron S. Scholes in 1993 for their pioneering work in the 1970s on pricing financial derivatives. In particular, they, along with Fischer Black, developed an analytical framework and simple mathematical formulae for pricing derivative assets, such as options and warrants, which have highly nonlinear payoff functions. Their work was the first step in the development of the derivatives industry and the whole risk management culture and practice in finance.

The close link between finance and nonlinear time series analysis is by no means accidental, being a consequence of four main factors. First, financial time series have always been considered ideal candidates for data-hungry nonlinear models. The fact that organized financial markets and information brokers (e. g., newspapers, data vendors, analysts, etc.) have been around for many years has meant that an abundance of high quality historical data exists. Most of this data is in the form of time series and usually spans several decades, sometimes exceeding a century. Furthermore, asset prices can now be collected at ultra-high frequencies, often less than a minute, so that sample sizes may run into millions of observations. Second, the poor forecasting performance of linear models allied to the prospect of obtaining large financial gains by 'beating the market' on the basis of superior forecasts produced by nonlinear time series models has provided a natural motive for researchers from several disciplines. Third, developments in the natural sciences since the 1980s with respect to chaos theory, nonlinear dynamics and complexity have fueled a 'nonlinearist' movement in finance and have motivated a new research agenda on relevant theories, models and testing procedures for financial time series. Underlying this movement was the concern that the apparent unpredictability of financial time series may simply be due to the inadequacy of standard linear models. Moreover, it was also thought that the irregular fluctuations in financial markets may not be the result of propagated exogenous random shocks but, rather, the outcome of some, hopefully low-dimensional, chaotic system (see the entry by Shintani on ► Financial Forecasting, Sensitive Dependence). Fourth, and most importantly, although the bulk of financial theory and practice is built upon affine models, a wealth of theoretical models and supporting empirical evidence has been published suggesting that the nature of some financial problems may be inherently nonlinear. Two prime examples are the time-varying and asymmetric nature of financial risk and the highly nonlinear relationships that arise in situations involving financial options and other derivatives.

Introduction

Traditionally, theorists and empirical researchers in finance and economics have had rather different views concerning nonlinearity. Theorists have shown some interest in nonlinearities and have used them in a variety of different ways, such as first order conditions, multimodality, ceilings and floors, regime switching, multiple equilibria, peso problems, bandwagon effects, bubbles, prey-predator dynamics, time-varying parameters, asymmetries, discontinuities and jump behavior, non-additivity, non-transitivity, etc. Theories and structural models that have nonlinear elements can be found in most areas of finance and economics (selective reviews with a focus mainly on economics are given by Brock and de Lima [19], Lorenz (see Chaps. 1–3 and 6 in [53]), Mullineux and Peng [67], Rosser [74]; other sources include Chap. 3 and pp. 114–147 in [40], [75]). Prominent examples include the noise-trader models of exchange rate determination [34,35], the target-zone exchange rate models [36,50] and the imperfect knowledge models [37]. Nonlinearities find their natural place in the theory of financial derivatives (for overviews, see [46,62]) and real options [26], where payoff functions and relationships between pricing variables are inherently highly nonlinear. The popularity of nonlinearities is limited by the prevailing equilibrium theory assumptions (convexity and continuity conditions, concavity of utility and production functions, constant returns to scale, intertemporally independent tastes and technology, rational aggregate expectations and behavior, etc.) which invariably lead to linear relationships.

For many years, nonlinearities were not a serious consideration when attempting to build empirical models. Alfred Marshall, one of the great pioneers of mathematical economics, epitomized the culture against nonlinear models when saying that "*natura non facit saltum*", or nature dislikes jumps. Although he contemplated the possibility of multiple equilibria and switching behavior and understood that this situation would entail a tendency for stable and unstable equilibria to alternate, he dismissed it as deriving "*from the sport of imagination rather than the observation of facts*". Correspondingly, in empirical and theoretical finance the mainstream approach has been to transform any nonlinearities to linearized forms using Taylor series expansions which excluded second-and higher-order terms. Since the 1990s, however, there has been a significant turn in favor of nonlinear modeling in finance. In addition to the reasons advanced earlier, this development has also been the result of advances in econometric estimation and of the widespread availability of cheap computer power. Some of the basic nonlinear models and relevant theories that have been used in finance will be discussed in the subsequent section (for a comprehensive review of the linear and nonlinear time series models used in finance, see [22,63]).

Basic Nonlinear Financial Time Series Models

Most of the theoretical and empirical research in financial economics has typically hypothesized that asset price time series are unit root stochastic processes with returns that are serially unpredictable. For many years it was thought that this unpredictability was necessary in order to ensure that financial markets function properly according to the Efficient Market Hypothesis (EMH; see the reviews by Fama [32,33]). Within this framework, a market is considered efficient with respect to a specific information set, an asset pricing model and a data generating process, respectively. For example, a general condition of efficiency is that market prices fully, correctly and instantaneously reflect the available information set. This is sometimes formalized as the Random Walk Hypothesis (RWH), which predicts that prices follow random walks with price changes that are unforecastable on the basis of past price changes. An even milder condition is that trading on the information set does not allow profits to be made at a level of risk that is inconsistent with the underlying asset pricing model. Although initially the EMH and RWH were thought to be an unavoidable consequence of the widely accepted paradigm of rational expectations, this was later refuted by a series of studies showing that random walk behavior was neither a necessary nor sufficient condition for rationally determined financial prices. Market efficiency has profound practical economic implications insofar as financial prices serve both as ways of integrating and distributing available information and as asset allocation devices.

One of the simplest models of financial prices that can be derived on the basis of unpredictability is the martingale process:

$$p_t = p_{t-1} + \varepsilon_t \tag{1}$$

where p_t is the price of an asset observed at time t and ε_t is the martingale increment or martingale difference. The martingale has the following properties: a) $E(|p_t|) < \infty$ for each t, b) $E(p_t|\Im_s) = p_s$ whenever $s \leqslant t$, where $\Im_s$ is the σ-algebra comprized of events determined by observations over the interval $[0, t]$, so that $\Im_s \subseteq \Im_t$ when $s \leqslant t$. The martingale possesses the Markov property since the differences $\Delta p_t = p_t - p_{t-1} = \varepsilon_t$ are unpredictable on the basis of past differences. By successive backward substitution in (1) we can express the current price as the accumulation of all past errors. In financial terms, errors can be thought to be the result of unexpected information or news. By restricting the differences ε_t to be identically and independently distributed (iid) we obtain what is often called the random walk process. The random walk is a term and assumption which is widely employed in finance. It was first used by Karl Pearson in a letter to Nature in 1905 trying to describe a mosquito infestation in a forest. Soon after, Pearson compared the process to the walk of an intoxicated man, hence the graphical term "drunkard's walk“.

By representing the random walk in continuous time with a growth rate μ, as is often useful when dealing with derivatives, we obtain the generalized Wiener process (also called Brownian motion or diffusion):

$$\mathrm{d}p_t = \mu\mathrm{d}t + \sigma\mathrm{d}w_t \tag{2}$$

where $\mathrm{d}w_t$ is a standard normal random variable. The parameters μ and σ are referred to in finance as the drift and volatility of the process, respectively. Another point worth mentioning is that in both discrete and continuous time the analysis is typically undertaken using logarithmically transformed prices. This precludes the paradoxical possibility of obtaining negative prices while, at the same time, regularizing the statistical behavior of the data. Assuming that prices are lognormally distributed means that logarithmic returns are normally distributed and can be calculated as $\log p_t - \log p_{t-1}$ or $\log(p_t/p_{t-1})$. These represent continuously compounded returns and are approximately equal to simple percentage returns.

Random walks, along with continuous-time mathematical finance, were formally introduced in 1900 by Louis Bachelier in his brilliant doctoral dissertation *Théorie de la Spéculation*. Under the supervision of the great Henri Poincaré, who first realized the possibility of chaotic motion, Bachelier developed the mathematical framework of random walks in continuous time in order to describe the unpredictable evolution of stock prices and to build the first option pricing model (biographical details of Bachelier are given in [58]). Random walks were independently discovered by Albert Einstein in 1905 and, of course, have since played a fundamental role in physics and mathematics. They were later rigorously treated, along with forecasting and nonlinear modeling, by Norbert Wiener, the father of cybernetics. Several important deviations from the Bachelierian random walk and normal distribution paradigm were developed several decades later by Benoit Mandelbrot and his co-authors (for an overview see [59], and the references given therein). This research developed around the generalized Central Limit Theorem (CLT), the stable family of distributions, long-term dependence processes, scaling and fractals. Indeed, it is clear that Mandelbrot views his research as similar to that of Bachelier in that both were inspired by finance and both found great applications later in physics or, to use Mandelbrot's words, both were cases of the "unexpected historical primacy of financial economics over physics" (see p. 174 in [59]).

Much of the motivation behind nonlinear time series modeling in finance has to do with certain empirical characteristics, or stylized facts, which have been observed over the years across many financial assets, markets and time periods. Since these characteristics were not always consistent with a linear data generating process, nonlinear models seemed to be a reasonable explanation. In particular, starting with Mandelbrot and others in the 1960s, several empirical studies have reported that financial assets typically have daily returns exhibiting:

- Nonnormality: skewed and leptokurtic (fat-tailed and high-peaked) unconditional distributions.
- Jump behavior: discontinuous variations that result in extreme observations.
- Volatility clustering: large (small) returns in magnitude tend to be followed by large (small) returns of either sign.
- Unpredictability: zero or weak serial autocorrelations in returns.

In order to illustrate these characteristics, we investigate the empirical behavior of daily logarithmic prices and returns (simply referred to as prices and returns hereafter) for the S&P 500 index. The series is publicly available from *Yahoo Finance*. The empirical analysis is undertaken using the econometric software packages *EViews* 5.0 by Quantitative Micro Software and *Time Series Modelling* 4.18 by James Davidson, respectively. The sample consists of 14,582 closing (end of the day) prices covering the period 3/1/1950–14/12/2007. The index is calculated as a weighted average of the common stock prices for the 500 largest firms traded on the New York Stock Exchange (NYSE) and is adjusted for dividends and splits. The S&P 500 is often used as a proxy for the so-called market portfolio and as a measure of the overall performance of the US stock market.

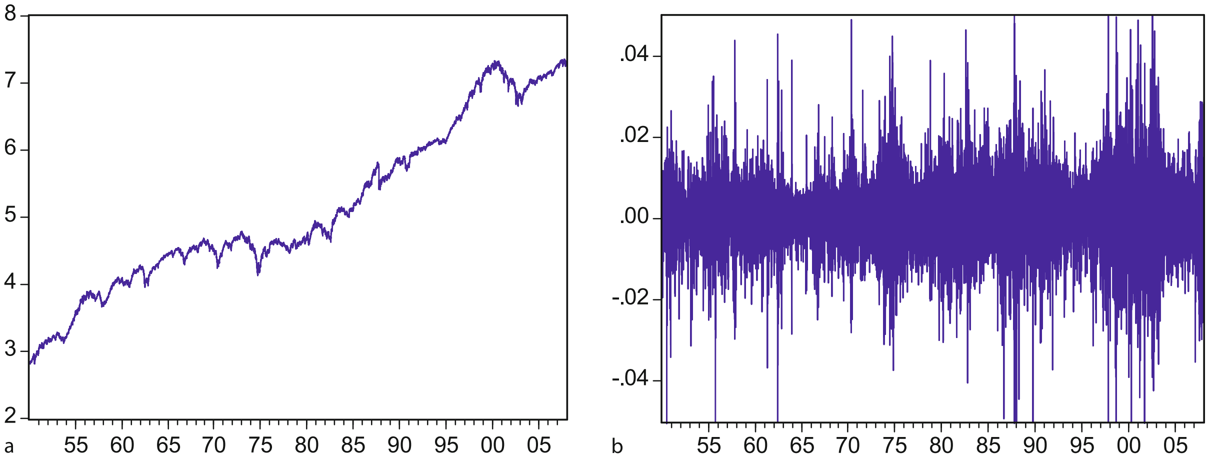
The prices depicted in the left part of Fig. 1 exhibit the upward drifting random walk behavior which is so routinely observed in financial time series. This is consistent with the fact that the series could not be predicted on the basis of past values using a member of the ARIMA class of linear models (as popularized by Box and Jenkins [16]). More specifically, the best fit was offered by an ARIMA(1,1,1) model, although with a disappointingly low *R*-squared statistic of just 0.64% (absolute *t*-statistics in brackets):

$$\Delta p_t = \underset{(3.9175)}{0.0003} - \underset{(3.3030)}{0.3013}\,\Delta p_{t-1} + \underset{(4.2664)}{0.3779}\,\varepsilon_{t-1} + \varepsilon_t\,. \quad (3)$$

Such weak linear serial predictabilities are often found at high sampling frequencies and are usually explained by market microstructures. They do not represent true predictabilities but, rather, result from specific market mechanisms and trading systems (see the survey by Biais et al. [12]).

A close examination of the return series, presented in the right part of Fig. 1, suggests the presence of large, discontinuous variations. In particular, we can count 26 daily returns which, in absolute value, exceed five standard deviations. This implies that such extreme events occur with a probability of 0.18% or, on average, almost once every two years (assuming 250 trading days in each calendar year). Under a normal distribution, such 'five-standard deviation' events should be extremely rare, with a probability of occurrence of only 0.00003%, or less than 3 days in every 40,000 years! The fat tails of the return distribution are also reflected in the kurtosis coefficient of 37.3, which is much larger than the value of 3 that corresponds to a normal distribution. In terms of asymmetry, the distribution is skewed to the left with the relevant coefficient estimated at −1.3.

Clearly, the normal distribution provides a poor approximation of reality here: the distribution of the errors in the random process described by (1) should be allowed to follow some non-Gaussian, fat-tailed and possi-



Financial Economics, Non-linear Time Series in, Figure 1
Daily S&P 500 log Index Prices *(left)* and Returns *(right)* (3/1/1950–14/12/2007) (Returns are trimmed to ±5% in order to improve the readability of the graph)

bly skewed distribution (see the entry by Haas and Pigorsch on ► Financial Economics, Fat-Tailed Distributions). Various distributions having these properties have been proposed, including the Student-*t*, the mixture of normals, double Weibull, generalized beta, Tukey's $g \times h$, generalized exponential, asymmetric scale gamma, etc. (see [48,71]). Although some of the non-Gaussian distributions that have been proposed have many desirable properties, empirical evidence regarding their appropriateness for describing financial returns has been inconclusive. Moreover, these distributions often bring with them acute mathematical problems in terms of representation, tractability, estimation, mixing and asymptotics. A distribution that has received considerable attention is the stable family (also known as the stable Paretian, Pareto–Lévy, or Lévy flight), which was initially proposed by Mandelbrot [54,55]. Stable distributions are highly flexible, have the normal and Cauchy as special cases, and can represent 'problematic' empirical densities that exhibit asymmetry and leptokurtosis. Furthermore, they are consistent with stochastic behavior that is characterized by discontinuities or jumps. From a theoretical point of view, stable distributions are particularly appealing since they are the limiting class of distributions in the generalized CLT, which applies to scaled sums of iid random variables with infinite variances. Stable distributions also exhibit invariance under addition, a property that is important for financial data, which are usually produced as the result of time aggregation. For a comprehensive discussion of these distributions, see [54,55,64,65,71,76].

In terms of the conditional distribution, it is evident from the graph of returns that the variance is not homogeneous across time, as one would expect for an iid process. In line with this observation, the autocorrelation of squared or absolute returns suggest the presence of strong dependencies in higher moments, something that in turn is indicative of conditional heteroskedasticity (see Fig. 3 below). On the basis of the above, it appears that the simple random walk model is far too restrictive and that the more general martingale process provides a better approximation to the data. Unlike the random walk, the martingale rules out any dependence in the conditional expectation of Δp_{t+1} on the information available at t, while allowing dependencies involving higher conditional moments of Δp_{t+1}. This property of martingales is very useful for explaining clusters in variances, since it allows persistence (correlation) in the conditional variances of returns.

It should be noted that much of the empirical work on nonlinear financial time series has involved modeling time varying variances. This concentration on the variance stems from it being the most widely used measure of risk, which, in orthodox finance theory, is the sole determinant of the expected return of any asset. Knowing the expected return enables the opportunity cost of any investment or asset to be estimated and, ultimately, to have a fair price put on its value by discounting all future revenues against the expected return. Variance was introduced in the path-breaking research of Nobel Laureate Harry Markowitz in the 1950s on investment portfolio selection, which laid the basis for what is known today as modern portfolio theory. The main innovation of Markowitz was that he treated portfolio selection as a tractable, purely quantitative problem of utility maximization under uncertainty, hence the term 'quant analysis'. Markowitz assumed that economic agents face a choice over two-dimensional indifference curves of investment preferences for risk and return. Under some additional assumptions, he obtained a solution to this problem and described the preferences of homogeneous investors in a normative manner using the mean and variance of the probability distribution of single period returns: such investors should optimize their portfolios on the basis of a 'mean-variance' efficiency criterion, which yields the investment with the highest expected return for a given level of return variance.

Let us now turn to some of the processes that have been used to model regularities in variance. For example, consider the GARCH(1,1) process, which has become very popular for modeling the conditional variance, σ_t^2, as a deterministic function of lagged variances and squared errors (see the entry by Hafner on ► GARCH Modeling):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{4}$$

where the ε_t are, in general, the residuals from a fitted conditional mean equation. This specification corresponds to a single-lagged version of the GARCH(p,q) (Generalized Autoregressive Conditional Heteroskedasticity) model proposed by Bollerslev [13] and can easily be modified to include additional lagged squared errors and variances. The GARCH model is an extension of the ARCH process originally proposed by Engle [27] and has served as the basis for the development of an extensive family of related models. For a review of this huge literature see, among others, [10,14,15,52,79], and Chap. 5 in [63]. Multivariate extensions of GARCH processes have also been proposed, but bring several computational and estimation problems (see [9,21]).

Two alternative approaches to modeling conditional variances in finance are extreme value estimators (see [17]) and realized variance (see [8]). Extreme value estimators depend on opening, closing, high and low prices during the trading day. Although they perform relatively well in terms of efficiency and are easy to estimate, they are quite badly biased. Realized variances are considered to be very accurate and are easily estimated as the sum of squared returns within a fixed time interval. Their limitation is that they require high frequency data at the intradaily level, which can be strongly affected by market microstructures and may not always be readily available. Rather than focusing on just the conditional variance, models have also been proposed for higher moments, such as conditional skewness and kurtosis (e. g., [43,47]).

To illustrate the application of some of the most popular GARCH parameterizations, consider again the S&P 500 return series. Using Maximum Likelihood (ML) estimation with t-student errors, the following 'GARCH(1,1)-in-Mean' (GARCH-M) model was obtained (absolute z-statistics appear in brackets):

$$\begin{aligned}
\Delta p_t &= \underset{(10.0994)}{0.0785}\,\sigma_t + \varepsilon_t \\
\sigma_t^2 &= \underset{(6.5616)}{5.77 \cdot 10^{-7}} + \underset{(16.3924)}{0.0684}\,\varepsilon_{t-1}^2 + \underset{(218.0935)}{0.9259}\,\sigma_{t-1}^2 \,.
\end{aligned}$$

In this model, originally proposed by Engle et al. [30], returns are positively related to the conditional standard deviation, σ_t. This is a particularly useful specification since it is directly consistent with Markowitz's theory about the positive relation between expected return and risk. In particular, the slope coefficient in the conditional mean equation can be interpreted as a relative risk aversion parameter, measuring how investors are compensated by higher returns for bearing higher levels of risk.

It is instructive to show in Fig. 2 both the estimated GARCH(1,1)-M conditional standard deviations and the standardized residuals, ε_t/σ_t. On the one hand, the model clearly produces mean reversion in volatility, which resembles the empirical behavior observed in the original series. Although the estimated conditional variance process appears to be highly persistent, it is nevertheless stationary since the sufficiency condition is satisfied because $\alpha_1 + \beta_1 = 0.0684 + 0.9259 = 0.9943 < 1$. On the other hand, the standardized residuals have a far more homogeneous conditional volatility than the original series and more closely resemble a white noise process. Moreover, the standardized residuals are closer to a normal distribution, with a kurtosis coefficient of 7.7, almost five times smaller than that of the original return series.

Careful inspection of the relationship between returns and conditional variance often reveals an asymmetric relationship. Threshold GARCH (TGARCH) and Exponential GARCH (EGARCH) are two of the specifications often used to model this commonly encountered nonlinearity. These models were estimated using the ML approach and the following conditional variance specifications were obtained.

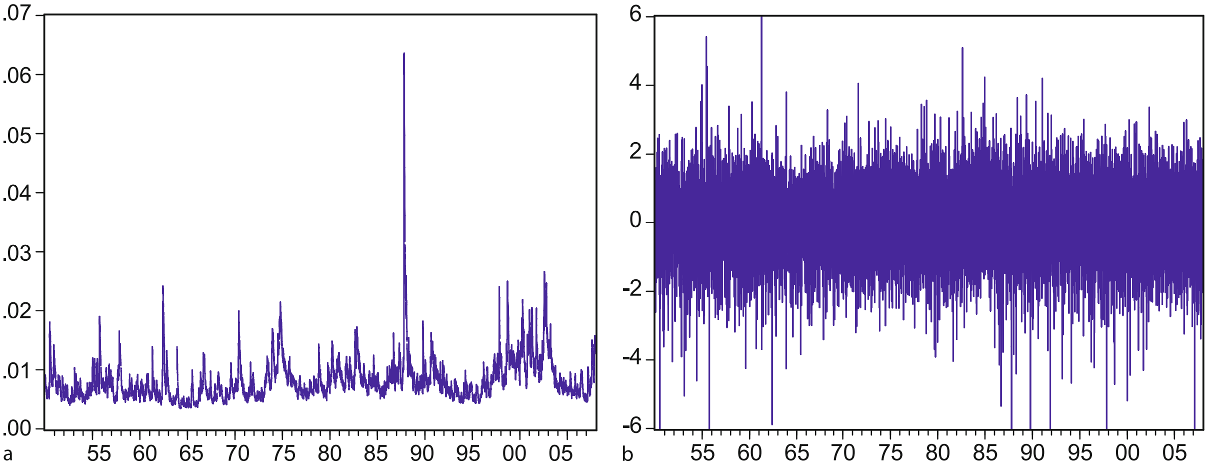
TGARCH

$$\begin{aligned}
\sigma_t^2 &= \underset{(8.5110)}{7.50 \cdot 10^{-7}} + \underset{(6.3766)}{0.0259}\,\varepsilon_{t-1}^2 \\
&\quad + \underset{(13.1058)}{0.0865}\,\varepsilon_{t-1}^2\, g + \underset{(224.9279)}{0.9243}\,\sigma_{t-1}^2
\end{aligned}$$

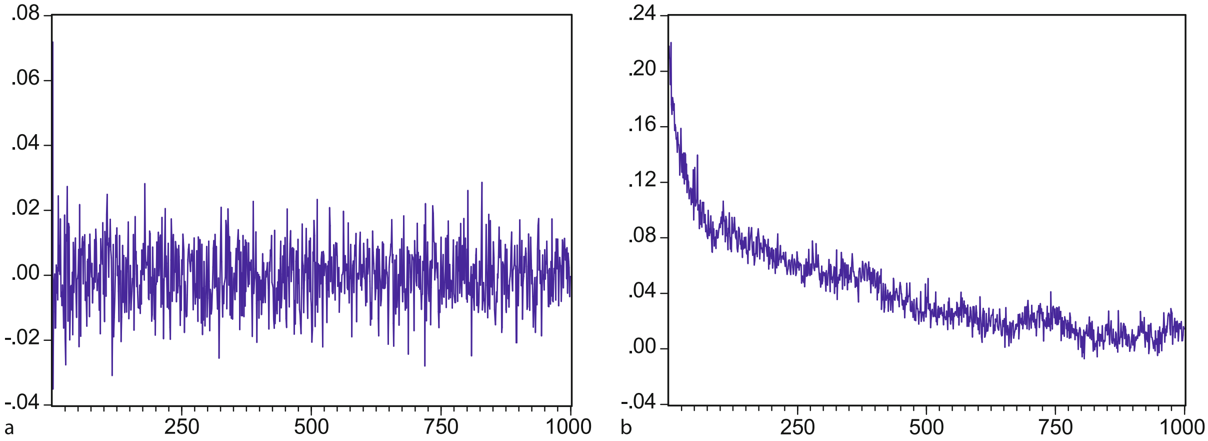
EGARCH

$$\begin{aligned}
\log(\sigma_t^2) &= -\underset{(13.7858)}{0.2221} + \underset{(16.9612)}{0.1209}\,|\varepsilon_{t-1}/\sigma_{t-1}| \\
&\quad - \underset{(15.9599)}{0.0690}\,\varepsilon_{t-1}/\sigma_{t-1} + \underset{(703.7161)}{0.9864}\,\log(\sigma_{t-1}^2)\,.
\end{aligned}$$

In the TGARCH model, the threshold parameter is defined as $g = 1$ if $\varepsilon_{t-1} < 0$ and 0 otherwise. Standard



Financial Economics, Non-linear Time Series in, Figure 2
GARCH(1,1)-M standard deviations *(left)* and standardized residuals *(right)* (Residuals are trimmed to ±6 standard deviations in order to improve the readability of the graph)



Financial Economics, Non-linear Time Series in, Figure 3
Autocorrelation function of S&P 500 simple *(left)* and absolute returns *(right)*

GARCH models, such as the GARCH-M estimated previously, assume that positive and negative errors (or news) have a symmetric effect on volatility. In the TGARCH and EGARCH models, news has an asymmetric effect on volatility depending on its sign. Specifically, in the TGARCH model news will have differential impacts on volatility depending on the signs and sizes of the coefficients on ε^2_{t-1} and $\varepsilon^2_{t-1} \cdot g$: good news ($\varepsilon_{t-1} > 0$) has an impact of 0.0259, while bad news ($\varepsilon_{t-1} < 0$) has a stronger impact of $0.0259 + 0.0865 = 0.1124$. Since the coefficient of $\varepsilon^2_{t-1} \cdot g$ is positive (0.0865), bad news tends to increase volatility, producing what is known as the 'leverage' effect. This was first observed in the 1970s and postulates that negative returns will usually reduce the stock price and market value of the firm, which in turn means an increase in leverage, i. e. a higher debt to equity ratio, and ultimately an increase in volatility. In the EGARCH model, forecasts are guaranteed to be positive since logarithms of the conditional variance are modeled. Since the sign of the coefficient on $\varepsilon_{t-1}/\sigma_{t-1}$ is non-zero and negative we can conclude that the effect of news on volatility is asymmetric and that a leverage effect is present.

Inspection of the autocorrelation functions (ACFs) in Fig. 3 for the returns and absolute returns of the S&P 500, the latter being a proxy for volatility, suggests very different behavior of the two series. While returns have an ACF that is typical of a white noise process, the autocorrelations of the absolute returns die out very slowly and become negative only after 798 lags! It turns out that many financial series have such extremely persistent or long-memory behavior. This phenomenon was first described by Mandelbrot [56,57] in the context of the 'Hurst effect' and was latter defined as fractional Brownian motion (see the relevant review by Brock [18]). Hosking [45] and Granger and Joyeux [39] modeled long-memory by extending the ARIMA class of processes to allow for fractional unit roots (for reviews, see [3,11,73,82]). The ARFIMA(p, d, q) model uses a fractional difference operator based on a binomial series expansion of the parameter d for any value between −0.5 and 0.5:

$$\Delta^d = 1 - dB + \frac{d\,(d-1)}{2!}B^2 - \frac{d\,(d-1)\,(d-2)}{3!}B^3 + \dots \tag{5}$$

where B is the backshift (or lag) operator with $B^m x_t = x_{t-m}$. In a similar fashion, investigating the existence of long-memory in the conditional variance of the returns could be undertaken in the context of a Fractional GARCH model (see [4]).

In our S&P 500 example, we have shown that nonlinearities enter through the conditional variance process and do so in an asymmetric manner. A natural question to ask is whether nonlinearities also exist in the conditional mean. Consider, for example, a generalization of the linear ARMA process

$$\Delta p_t = f(\Delta p_{t-i}, \varepsilon_{t-i}) + \varepsilon_t \tag{6}$$

where $f\,()$ is a nonlinear function and $\Delta p_{t-i}, \varepsilon_{t-i}$ are lagged price differences and errors, respectively. A wide variety of testing procedures have been proposed for examining the possibility of nonlinearities in the conditional mean process (for reviews see the relevant sections in [40], and [63]). Here we use the BDS test of the null hypothesis of serial independence, which has been widely applied and has been shown to have good power against a variety of nonlinear alternatives (see [20]). The test is inspired by chaos theory and phase space analysis and is based on the concept of the correlation integral. Specifically, the test relies on the property that, for an iid series, the probability of the distance between any two points being no greater than a predefined distance (ε) should be constant. A joint probability can also be calculated for sets comprising multiple pairs of points chosen by moving through consecutive sequences of observations in the sample. The number of consecutive data points used in such a set is called the (embedding) dimension and may be chosen by the user. Brock et al. [20] constructed an asymptotically normally distributed test statistic for the constancy of the distance ε between points. When this test was applied to the residuals from an MA(1)-EGARCH(1,1) model fitted to the S&P 500 returns, it was always insignificant across a variety of dimensions, implying that any nonlinear dependencies in the returns are due solely to GARCH effects.

An agnostic, yet often convenient, way to approximate the unknown nonlinear function (6) is to consider some nonparametric estimator (see [72]). Although several nonparametric estimators have been used with mixed success, one of the most popular is based on the neural network family of models (see [80]). A rich variety of parametric nonlinear functions have also been proposed in finance. A convenient and intuitive way of introducing nonlinearity is to allow 'regime switching' or 'time-variation' in the parameters of the data generating process (for a review see [70]). Three of the most popular approaches in this category are the Markov switching, the Threshold Autoregressive (TAR) and the Smooth Transition (STAR) models. In the first approach (for a popular implementation, see [41,42]), the model parameters switch according to a multiple (typically two) unobserved state Markov process. In TAR models (see [81], for a comprehensive description), nonlinearities are captured using piecewise autoregressive linear models over a number of different states. For example, consider the simple two regime case:

$$x_t = \begin{cases} \omega_1 + \sum_{i=1}^{p} \varphi_{1i} x_{t-i+1} + \sigma_1 \varepsilon_t\,, & s_{t-d} < c \\ \omega_2 + \sum_{i=1}^{p} \varphi_{2i} x_{t-i+1} + \sigma_2 \varepsilon_t\,, & s_{t-d} \geqslant c \end{cases} \tag{7}$$

where c is the threshold value, s_t is a threshold variable, d is a delay parameter assumed to be less than or equal to p, and the ε_t are iid standard normal variates assumed to be independent of lagged s_ts. The threshold variable is often determined by a linear combination of the lagged x_ts, in which case we obtain the Self Exciting TAR (SETAR) model. This has become a popular parameterization in finance since it can produce different dynamic behavior across regimes with characteristics such as asymmetry, limit cycles, jumps and time irreversibility (recall the TGARCH model introduced earlier, which has a related specification). STAR models allow a smooth switch between regimes using a smooth transition function. Transition functions that have been considered include the cumulative distribution of the standard normal, the exponential (ESTAR) and the logistic (LSTAR).

It is instructive to see how regime switching can be applied in the context of asset pricing models (for a comprehensive treatment of asset pricing, see [24]). The best known and most influential framework, which builds upon Markowitz's portfolio theory, is the Capital Asset Pricing Model (CAPM) proposed by Sharpe, Lintner, Black and others. The CAPM can be expressed as a single-period equilibrium model:

$$E(r_i) = r_\mathrm{f} + \beta_i \left[E(r_\mathrm{m}) - r_\mathrm{f}\right] \tag{8}$$

where $E(r_i)$ is the expected return on asset i, $E(r_\mathrm{m})$ is the expected return on the market portfolio, r_f is the risk-free interest rate, and the slope β_i is the so-called beta coefficient of asset i, measuring its systematic risk. Empirical implementations and tests of the CAPM are usually based on the 'excess market' and 'market model' regressions, respectively

$$r_{i,t} - r_{\mathrm{f},t} = r_{\mathrm{f},t} + \beta_i \left[r_{\mathrm{m},t} - r_{\mathrm{f},t}\right] + \varepsilon_{i,t} \tag{9}$$

and

$$r_{i,t} = \alpha_i + \beta_i r_{\mathrm{m},t} + \varepsilon_{i,t} \,. \tag{10}$$

The variance of the residuals $\varepsilon_{i,t}$ reflects the unsystematic risk in asset i. In practice the CAPM is typically estimated using ordinary least squares regression with five years of monthly data. A wealth of empirical evidence has been published showing that the basic assumptions of the CAPM regressions with respect to parameter stability and residual iid-ness are strongly refuted (see [60]). In particular, betas have been found to be persistent but unstable over time due to factors such as stock splits, business cycle conditions, market maturity and other political and economic events. In order to demonstrate the modeling of time-varying betas in the CAPM, consider first the simple market model regression for the stock returns of Tiffany & Co (listed on the New York Stock Exchange) against S&P 500 returns:

$$r_t = \underset{(17.5396)}{1.4081}\, r_{\mathrm{m},t} + \varepsilon_t , \qquad R^2 = 28.75\% \,.$$

The regression was estimated using weekly returns from 30/12/1987 to 14/12/2007, a total of 1,044 observations. The R^2 statistic denotes the proportion of total risk that can be explained by the model and which is thus systematic. The beta coefficient is significantly higher than unity, suggesting that the stock is 'aggressive' in that it carries more risk than the market portfolio. Allowing the beta coefficient to switch according to a Markov process produces the following two-regime market model:

$$r_t = \begin{cases} \text{Regime 1:}\ \underset{(1.9220)}{0.4797} r_{\mathrm{m},t} + \varepsilon_t \\ \text{Regime 2:}\ \underset{(10.6381)}{1.9434} r_{\mathrm{m},t} + \varepsilon_t \end{cases} \qquad R^2 = 41.63\% \,.$$

The explanatory power of the model has increased significantly and the stock is now characterized by both passive ($\beta = 0.4797 < 1$) and aggressive ($\beta = 1.9434 > 1$) systematic risk behavior regimes. The Markov transition probabilities $P\left(i \,|\, j\right)$, $j = 1, 2$, were estimated as $P(1|1) = 0.6833$, $P(1|2) = 0.3167$, $P(2|1) = 0.2122$ and $P(2|2) = 0.7878$. The smoothed probabilities for regime 1 are depicted in Fig. 4 and are seen to be rather volatile, so that the returns switch regimes rather frequently. For a discussion of the threshold CAPM see [2].

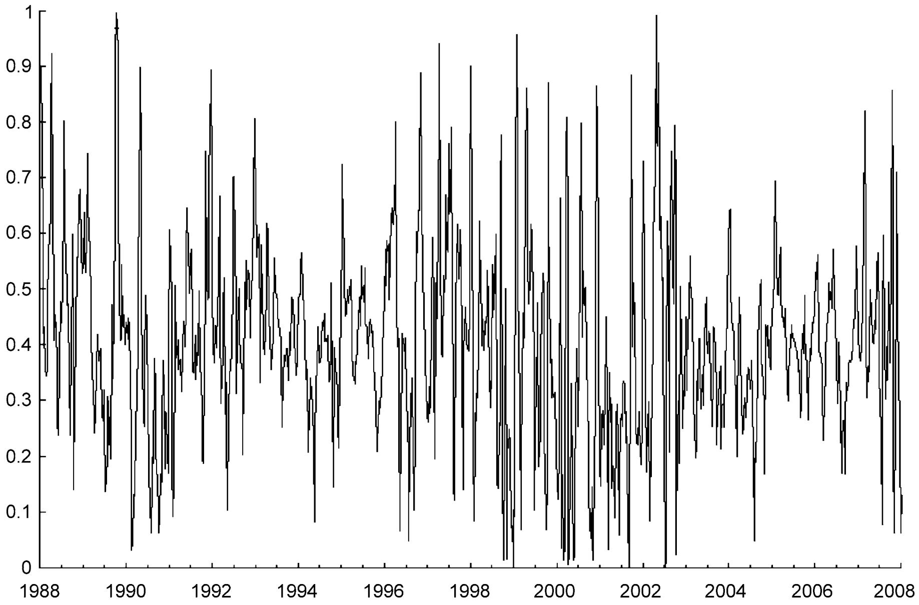
Another important category of models allows for nonlinear relationships between persistent financial time series. The most popular framework here is that of cointegration, which deals with variables that are individually nonstationary but have some joint stationary representation. For example, consider the linear combination of two unit root ($I(1)$) processes x_t and y_t

$$x_t = a + y_t + \varepsilon_t \,. \tag{11}$$

In general, ε_t will also be $I(1)$. However, as shown by Engle and Granger [29], if ε_t is actually $I(0)$, then x_t and y_t are said to be (linearly) cointegrated and will have an error-correction representation which, for example, could take the form

$$\Delta x_t = -\gamma \varepsilon_{t-1} + u_t \tag{12}$$

where $-\gamma$ denotes the strength of reversal to the equilibrium cointegrating relationship through the error-correction term, i. e., the lagged residual from the cointegrating regression (11). The finance literature has considered nonlinear generalizations of both the cointegrating regression (11) and the error-correction model (12) (see the entry by Escribano et al. on ► Econometrics: Non-linear Cointegration). Nonlinear error-correction mechanisms can be accommodated rather straightforwardly within the cointegration analysis framework, with the residuals from a linear cointegration relationship entering a nonlinear error-correction model. It has been shown that such nonlinearities may arise simply because of complex relationships between variables (see pp. 59–61 in [40]). Justifications in terms of finance theory have been based on factors such as arbitrage in the presence of transaction costs, heterogeneity among arbitrageurs, existence of partial adjustment models and market segmentation, agents' maximizing or minimizing behavior, constraints on central bank intervention, and intertemporal choice behavior under asymmetric adjustment costs. While almost all the different nonlinear specifications discussed previously have also been applied in error-correction modeling, threshold models hold a prominent position, as they allow large errors from equilibrium, i. e., those above some threshold, to be corrected while small errors are ignored (see, for example, [6]). The use of nonlinearities directly within the coin-



Financial Economics, Non-linear Time Series in, Figure 4
Tiffany stock Markov switching market model smoothed probabilities for Regime 1 of 2

tegrating relationship is not as straightfoward and brings several conceptual and estimation problems (see [63]).

Returning to the bivariate market model setting, it has been found that cointegrating relationships do exist between stock prices and index levels (see [60]). In our example, the logarithms of Tiffany's stock prices are cointegrated with S&P 500 logarithmic price levels. The following asymmetric error correction model was then estimated:

$$r_t = \underset{(3.0329)}{-0.0168}\,\varepsilon_{t-1}\,g + u_t$$

where g is the heavyside function defined previously with $g = 1$ if $\varepsilon_{t-1} < 0$ and 0 otherwise, ε_{t-1} being obtained from the cointegrating regression.

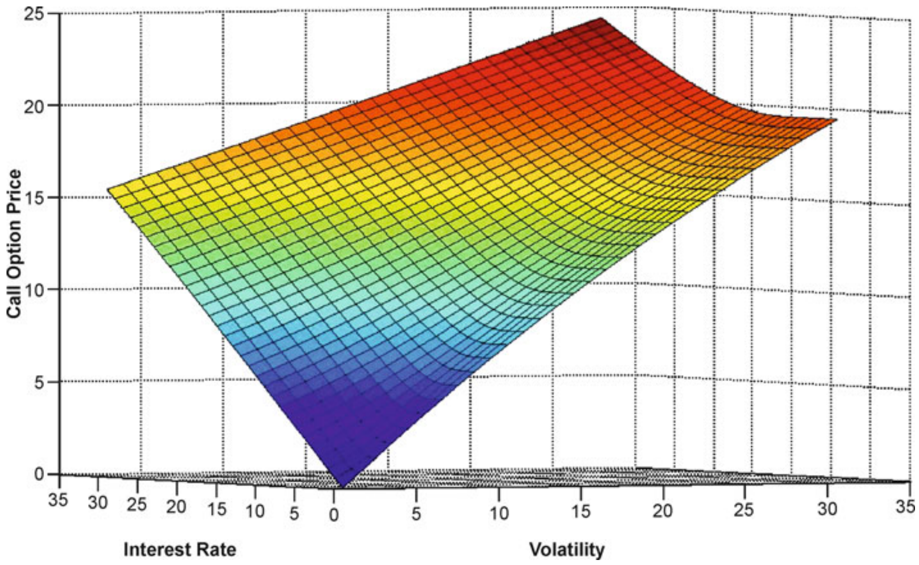
Several studies have shown that empirical characteristics and regularities, such as those discussed previously are very unlikely to remain stable if the sampling frequency of the data changes. For example, we find that if the S&P 500 returns are estimated at an annual frequency using the first available January price, then their distribution becomes approximately Gaussian with skewness and kurtosis coefficients estimated at −0.4 and 2.7, respectively. The annual prices are highly predictable using an ARIMA(2,1,2) process with an impressive adjusted R-squared value of 15.7%. Moreover, standard tests of heteroskedasticity suggest that the variance of annual returns can be assumed to be constant! In contrast, for very high sampling frequencies, say at the intradaily or tick-by-tick level, the data behave in a different manner and are characterized by strong seasonalities, e. g., variances and volumes follow an inverse J shape throughout the trading day (see the review by Goodhart and O'Hara [38], and the discussion in [28]).

Finally, let us now turn our discussion to models in a continuous time setting. As previously mentioned, the analysis of derivatives provides a natural setting for nonlinear modeling since it deals with the pricing of assets with highly nonlinear payoff functions. For example, under the widely used Black–Scholes option pricing model (see [46], for a thorough description), stock prices are lognormally distributed and follow a Wiener process. The Black–Scholes model allows for highly nonlinear relationships between the pricing variables and parameters, as shown in Fig. 5.

Another popular use of continuous time processes is in modeling the autonomous dynamics of processes such as interest rates and the prices of stocks and commodities. A generic stochastic differential equation that can be used to nest alternative models is the following:

$$\mathrm{d}S_t = \mu\,(S_t, t)\,\mathrm{d}t + \sigma\,(S_t, t)\,\mathrm{d}W_t + y\,(S_t, t)\,\mathrm{d}q_t \quad (13)$$

where S_t is the price at time t, $\mathrm{d}W_t$ is a standard Wiener process, $\mu\,(S_t, t)$ is the drift, and $\sigma\,(S_t, t)$ is the diffusion coefficient. Both the drift and diffusion coefficients are assumed to be functions of time and price, respectively. A jump component is also allowed by incorporat-



Financial Economics, Non-linear Time Series in, Figure 5
Call option prices, volatility and interest rate in the Black–Scholes model (Call option prices were estimated using the Black–Scholes model assuming a strike price of 50, 1 year time to maturity and a zero dividend yield)

ing a Poisson process, $\mathrm{d}q_t$, with a constant arrival parameter λ, i. e., $\Pr\{\mathrm{d}q_t = 1\} = \lambda \mathrm{d}t$ and $\Pr\{\mathrm{d}q_t = 0\} = 1 - \lambda \mathrm{d}t$: y is the jump amplitude, also a function of time and price. $\mathrm{d}W_t$, $\mathrm{d}q_t$ and y are assumed to be mutually independent processes. Several nonlinear models can be obtained by combining various assumptions for the components $\mu(S_t, t)$, $\sigma(S_t, t)$ and $y(S_t, t)$. For example, consider the following processes.

Mean Reverting Square-Root Process (MRSRP)

$$\mathrm{d}S_t = \kappa(\theta - S_t)\,\mathrm{d}t + \sigma\sqrt{S_t}\mathrm{d}W_t \tag{14}$$

Constant Elasticity of Variance (CEV)

$$\mathrm{d}S_t = \kappa(\theta - S_t)\,\mathrm{d}t + \sigma S_t^{\gamma}\,\mathrm{d}W_t \tag{15}$$

Geometric Wiener Process augmented by Jumps (GWPJ)

$$\mathrm{d}S_t = \left(\mu - \lambda\mu_j\right) S_t \mathrm{d}t + \sigma S_t \mathrm{d}W_t + \left(\mathrm{e}^{y} - 1\right) S_t \mathrm{d}q_t \tag{16}$$

MRSRP augmented by Jumps (MRSRPJ)

$$\mathrm{d}S_t = \kappa(\theta - S_t)\,\mathrm{d}t + \sigma\sqrt{S_t}\mathrm{d}W_t + y\mathrm{d}q_t\,. \tag{17}$$

Model (14) has been widely used in modeling interest rates (e. g., [1,23,25]) and stochastic volatility (e. g., [44,68]). Process (16) is often used for representing the dynamics of stock prices and indices (e. g., [61]). Model (17) has been recently employed by several researchers for modeling volatility, because it allows rapid changes in volatility during times of market stress (e. g., [31]). While process (16) has a proportional structure, with μ being the expected return of the asset per unit of time and σ its volatility, the other processes have mean reverting drifts. In Eqs. (14), (15) and (17) κ is the speed of mean reversion, θ is the unconditional long-run mean, and σ the volatility of the price process. In Eq. (15), γ is a free parameter to be estimated that determines the dependence of the diffusion component on the current level of S. In Eqs. (16) and (17), λ is the average number of jumps per year and y is the jump size, which can be drawn from a normal or a double exponential distribution (see [49]).

An alternative way of representing the conditional variance is to use a stochastic volatility model, in which volatility is driven by its own noise (see the entry by Andersen and Benzoni on ► Stochastic Volatility). Stochastic volatility models are advantageous in that they are very flexible and have representations in both discrete and continuous time. The square root volatility model (also known as a scalar affine diffusion), proposed by Heston [44], is one of the most popular models in this area and is represented by the stochastic processes

$$\begin{aligned} \mathrm{d}\log(p_t) &= (\mu - 0.5\sigma_t)\,\mathrm{d}t + \sqrt{V_t}\mathrm{d}W_{1t} \\ \mathrm{d}V_t &= (\alpha - \beta\sigma_t)\,\mathrm{d}t + \sigma_V\sqrt{V_t}\mathrm{d}W_{2t} \end{aligned} \tag{18}$$

where V_t is the instantaneous (latent) stochastic volatility, which is assumed to follow a mean reverting square root process. The parameter k measures the speed of mean reversion, while θ is the unconditional long run mean. $\mathrm{d}W_{1t}$ and $\mathrm{d}W_{2t}$ are Brownian motions with instantaneous correlation $\rho \mathrm{d}t$.

Future Directions

The coverage in this essay has, unavoidably, been far from exhaustive. The realm of relevant nonlinear models and theories in finance is extremely rich and is developing fast (a useful review of new developments is [66]). By transcending the representative agent framework and by extending the standard notion of rationality, researchers are now allowing for interactions between heterogeneous groups of investors using agent based models (for an overview of these fascinating developments, see [51] and the entry on ► Finance, Agent Based Modeling in by Manzan). While such approaches can reproduce stylized facts such as volatility clustering and long-term dependencies, it remains to be seen how they can be standardized and applied to the solution of specific problems by academics and practitioners.

Bibliography

1. Aït-Sahalia Y (1999) Transition densities for interest rate and other nonlinear diffusions. J Finance 54:1361–1395
2. Akdeniz L, Altay-Salih A, Caner M (2003) Time varying betas help in asset pricing: The threshold CAPM. Stud Nonlinear Dyn Econom 6:1–16
3. Baillie RT (1996) Long memory processes and fractional integration in econometrics. J Econom 73:5–59
4. Baillie RT, Bollerslev T, Mikkelson HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. J Econom 74:3–30
5. Bakshi G, Ju N, Yang H (2006) Estimation of continuous-time models with an application to equity volatility dynamics. J Financ Econom 82:227–249
6. Balke NS, Fomby TB (1997) Threshold cointegration. Int Econom Rev 38:627–645
7. Barkley Rosser J Jr (1999) On the complexities of complex economic dynamics. J Econom Perspect 13:169–192
8. Barndorff-Nielsen OE, Graversen SE, Shephard N (2004) Power variation and stochastic volatility: A review and some new results. J Appl Probab 41:133–143
9. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109
10. Bera AK, Higgins ML (1993) On ARCH models: Properties, estimation and testing. J Econom Surv 7:305–366
11. Beran JA (1992) Statistical methods for data with long-range dependence. Stat Sci 7:404–427
12. Biais B, Glosten L, Spatt C (2005) Market microstructure: A survey of microfoundations, empirical results, and policy implications. J Financ Mark 8:217–264
13. Bollerslev T (1986) Generalised autoregressive conditional heteroskedasticity. J Econom 31:307–27
14. Bollerslev T, Chou RY, Kroner KF (1992) ARCH modelling in finance: A review of the theory and empirical evidence. J Econom 52:5–59
15. Bollerslev T, Engle RF, Nelson DB (1994) ARCH Models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. New York, North-Holland, pp 2959–3038
16. Box GEP, Jenkins GM (1976) Time Series Analysis: Forecasting and Control. Rev. Edn., Holden Day, San Francisco
17. Brandt MW, Diebold FX (2006) A no-arbitrage approach to range-based estimation of return covariances and correlations. J Bus 79:61–74
18. Brock WA (1999) Scaling in economics: A reader's guide. Ind Corp Change 8:409–446
19. Brock WA, de Lima PJF (1996) Nonlinear time series, complexity theory, and finance. In: Maddala GS, Rao RS (eds) Handbook of Statistics, vol 14. Elsevier, Amsterdam, pp 317–361
20. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econom Rev 15:197–235
21. Brooks C (2006) Multivariate Volatility Models. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1. Econometric Theory. Palgrave Macmillan, Basingstoke, pp 765–783
22. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, New Jersey
23. Chan KC, Karolyi A, Longstaff FA, Sanders AB (1992) An empirical comparison of alternative models of the short-term interest rate. J Finance 47:1209–1227
24. Cochrane JH (2005) Asset Pricing. Princeton University Press, Princeton
25. Cox JC, Ingersoll JE, Ross SA (1985) A theory of the term structure of interest rates. Econometrica 53:385–408
26. Dixit AK, Pindyck RS (1994) Investment under Uncertainty. Princeton University Press, Princeton
27. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50:987–1008
28. Engle RF (2000) The econometrics of ultra-high frequency data. Econometrica 68:1–22
29. Engle RF, Granger CWJ (1987) Cointegration and error correction: representation, estimation and testing. Econometrica 55:251–276
30. Engle RF, Lilien DM, Robins RP (1987) Estimating time varying risk premia in the term structure: the ARCH-M model. Econometrica 55:391–408
31. Eraker B, Johannes M, Polson N (2003) The impact of jumps in volatility and returns. J Finance 53:1269–1300
32. Fama EF (1991) Efficient capital markets, vol II. J Finance 26:1575–1617
33. Fama EF (1998) Market efficiency, long-term returns, and behavioural finance. J Financ Econom 49:283–306
34. Frankel FA, Froot KA (1987) Using survey data to test propositions regarding exchange rate expectations. Am Econom Rev 77:33–153
35. Frankel FA, Froot KA (1988) Chartists, fundamentalists and the demand for dollars. Greek Econom Rev 10:49–102
36. Froot KA, Obstfeld M (1991) Exchange-rate dynamics under stochastic regime shifts – A unified approach. J Int Econom 31:203–229

37. Goldberg MD, Frydman R (1996) Imperfect knowledge and behaviour in the foreign exchange market. Econom J 106: 869–893
38. Goodhart CAE, O'Hara M (1997) High frequency data in financial markets: Issues and applications. J Empir Finance 4:73–114
39. Granger CWJ, Joyeux R (1980) An introduction to long memory time series models and fractional differencing. J Time Ser Anal 1:15–29
40. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, New York
41. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
42. Hamilton JD (1990) Analysis of time series subject to changes in regime. J Econom 45:39–70
43. Hansen BE (1994) Autoregressive conditional density estimation. Int Econom Rev 35:705–730
44. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Financ Stud 6:327–343
45. Hosking JRM (1981) Fractional differencing. Biometrika 68:165–176
46. Hull JC (2005) Options, Futures and Other Derivatives, 6th edn. Prentice Hall, Upper Saddle River
47. Jondeau E, Rockinger M (2003) Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. J Econom Dyn Control 27:1699–1737
48. Kon S (1984) Models of stock returns – a comparison. J Finance 39:147–65
49. Kou SG (2002) A jump-diffusion model for option pricing. Management Sci 48:1086–1101
50. Krugman PR (1991) Target zones and exchange-rate dynamics. Q J Econom 106:669–682
51. LeBaron B (2006) Agent-based Computational Finance. In: Tesfatsion L, Judd K (eds) Handbook of Computational Economics. North-Holland, Amsterdam, pp 1187–1232
52. Li WK, Ling S, McAleer M (2002) Recent theoretical results for time series models with GARCH errors. J Econom Surv 16: 245–269
53. Lorenz HW (1989) Nonlinear Dynamical Economics and Chaotic Motion. Springer, New York
54. Mandelbrot BB (1963) New methods in statistical economics. J Political Econom 71:421–440
55. Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–419
56. Mandelbrot BB (1969) Long-run linearity, locally Gaussian process, H-spectra, and infinite variances. Int Econom Rev 10: 82–111
57. Mandelbrot BB (1972) Statistical methodology for nonperiodic cycles: From the covariance to R/S analysis. Ann Econom Soc Measurement 1/3:259–290
58. Mandelbrot BB (1989) Louis Bachelier. In: The New Palgrave: Finance. Macmillan, London, pp 86–88
59. Mandelbrot BB (1997) Three fractal models in finance: Discontinuity, concentration, risk. Econom Notes 26:171–211
60. Markellos RN, Mills TC (2003) Asset pricing dynamics. Eur J Finance 9:533–556
61. Merton RC (1976) Option prices when underlying stock returns are discontinuous. J Financ Econom 3:125–144
62. Merton RC (1998) Applications of option-pricing theory: Twenty-five years later. Am Econom Rev 88:323–347
63. Mills TC, Markellos RN (2008) The Econometric Modelling of Financial Time Series, 3rd edn. Cambridge University Press, Cambridge
64. Mittnik S, Rachev ST (1993) Modeling asset returns with alternative stable distributions. Econom Rev 12:261–330
65. Mittnik S, Rachev ST (1993) Reply to comments on "Modeling asset returns with alternative stable distributions" and some extensions. Econom Rev 12:347–389
66. Mizrach B (2008) Nonlinear Time Series Analysis. In: Blume L, Durlauf S (eds) The New Palgrave Dictionary of Economics, 2nd edn. Macmillan, London, pp 4611–4616
67. Mullineux A, Peng W (1993) Nonlinear business cycle modelling. J Econom Surv 7:41–83
68. Pan J (2002) The jump-risk premia implicit in options: Evidence from an integrated time-series study. J Financ Econom 63:3–50
69. Poitras G (2000) The Early History of Financial Economics. Edward Elgar, Cheltenham, pp 1478–1776
70. Potter S (1999) Nonlinear time series modelling: An introduction. J Econom Surv 13:505–528
71. Rachev ST, Menn C, Fabozzi FJ (2005) Fat Tailed and Skewed Asset Distributions. Wiley, New York
72. Racine JS, Ullah A (2006) Nonparametric Econometrics. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1: Econometric Theory. Palgrave Macmillan, Basingstoke, pp 1001–1034
73. Robinson PM (2003) Long Memory Time Series. In: Robinson PM (ed) Time Series with Long Memory. Oxford University Press, London, pp 4–32
74. Rosser JB Jr (1991) From Catastrophe to Chaos: A General Theory of Economic Discontinuities. Kluwer Academic, Norwell
75. Rostow WW (1993) Nonlinear Dynamics: Implications for economics in historical perspective. In: Day RH, Chen P (eds) Nonlinear Dynamics and Evolutionary Economics. Oxford University Press, Oxford
76. Samorodnitsky G, Taqqu MS (1994) Stable Non-Gaussian Random Processes. Chapman and Hall, New York
77. Schumpeter JA (1939) Business Cycles. McGraw-Hill, New York
78. Schwartz E (1997) The stochastic behavior of commodity prices: Implications for valuation and hedging. J Finance 52:923–973
79. Teräsvirta T (2006) An Introduction to Univariate GARCH Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, New York
80. Teräsvirta T, Medeiros MC, Rech G (2006) Building neural network models for time series: A statistical approach. J Forecast 25:49–75
81. Tong H (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford
82. Velasco C (2006) Semiparametric Estimation of Long-Memory Models. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1: Econometric Theory. Palgrave MacMillan, Basingstoke, pp 353–95

Financial Economics, Return Predictability and Market Efficiency

Stijn van Nieuwerburgh[1], Ralph S. J. Koijen[2]
[1] Department of Finance, Stern School of Business, New York University, New York, USA
[2] Department of Finance, Tilburg University, Tilburg, The Netherlands

Article Outline

Glossary

Stock return The stock return in this entry refers to the return on the portfolio of all stocks that are traded on the three largest equity markets in the US: the NYSE, NASDAQ, and AMEX. The return is measured as the price of the stock at the end of the year plus the dividends received during the year divided by the price at the beginning of the year. The return of each stock is weighted by its market capitalization when forming the portfolio. The source for the data is CRSP.

Dividend-price ratio and dividend yield The dividend-price ratio of a stock is the ratio of the dividends received during the year divided by the price of the stock at the *end* of the year. The dividend yield, instead, is the ratio of the dividends received during the year divided by the price of the stock at the *beginning* of the year. The stock return is the sum of the dividend yield and the capital gain yield, which measures the ratio of the end-of-year stock price to the beginning-of-year stock price.

Predictability A stock return r_{t+1} is said to be predictable by some variable x_t if the expected return conditional on x_t, $E[r_{t+1} \mid x_t]$, is different from the unconditional expected return, $E[r_{t+1}]$. No predictability means that the best predictor of tomorrow's return is the constant, unconditional average return, i. e., $E[r_{t+1} \mid x_t] = E[r_{t+1}]$. When stock returns are unpredictable, stock prices are said to follow a random walk.

Market model The market model links the return on any asset i, r_{it} to the return on the market portfolio (r_t). Under joint normality of returns, it holds:

$$r_{it} = \alpha_i + \beta_i r_t + \varepsilon_{it} , \qquad (1)$$

with $E[\varepsilon_{it}] = 0$ and $\text{Var}[\varepsilon_{it}] = \sigma^2_{\varepsilon_i}$, see [16]. The typical assumption in the literature until the 1980s has been that $E[r]$ is constant.

Definition of the Subject

The efficient market hypothesis, due to [21,22] and [23], states that financial markets are efficient with respect to a particular information set when prices aggregate all available information. Testing the efficient market hypothesis requires a "market model" which specifies how information is incorporated into asset prices. Efficiency of markets is then synonymous with the inability of investors to make economic, i. e., risk-adjusted, profits based on this information set [36]. The question of market efficiency and return predictability is of tremendous importance for investors and academics alike. For investors, the presence of return predictability would lead to different optimal asset allocation rules. Failing to make portfolios conditional on this information may lead to substantial welfare losses. For academics, return predictability or the lack thereof has substantial implications for general equilibrium models that are able to accurately describe the risks and returns in financial markets.

Introduction

Until the 1980s, the standard market model assumed constant expected returns. The first empirical evidence, which showed evidence that returns were predictable to some extent, was therefore interpreted as a sign of market inefficiency [25,54]. [56] proposed the alternative explanation of time-varying expected returns. This prompted the question of why aggregate stock market returns would be time varying in equilibrium. [23] provides a summary of this debate.

Recently developed general equilibrium models show that expected returns can indeed be time varying, even if markets are efficient. Time-variation in expected returns can result from time-varying risk aversion [11], long-run consumption risk [5], or time-variation in risk-sharing opportunities, captured by variation in housing collateral [44]. Predictability of stock returns is now, by-and-large, interpreted as evidence of time-varying expected returns rather than market inefficiency.

Motivating Predictive Regressions

Define the gross return on an equity investment between period t and period $t + 1$ as

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} \, ,$$

where P denotes the stock price and D denotes the dividend. [9] log-linearizes the definition of a return to obtain:

$$r_{t+1} = k + \Delta d_{t+1} + \rho dp_{t+1} - dp_t \, . \tag{2}$$

All lower-case letters denote variables in logs; d_t stands for dividends, p_t stands for the price, $dp_t \equiv d_t - p_t$ is the log dividend–price ratio, and r_t stands for the return. The constants k and $\rho = (1 + \exp(\overline{dp}))^{-1}$ are related to the long-run average log dividend–price ratio $\overline{dp}$. By iterating forward on Eq. (2) and by imposing a transversality condition (i. e., we rule out rational bubbles), one obtains

$$dp_t = \overline{dp} + E_t \sum_{j=1}^{\infty} \rho^{j-1} \left[(r_{t+j} - \bar{r}) - (\Delta d_{t+j} - \bar{d}) \right] . \tag{3}$$

Since this equation holds both ex-post and ex-ante, an expectation operator can be added on the right-hand side. This equation is one of the central tenets of the return predictability literature, the so-called Campbell and Shiller [12,13] equation. It says that, as long as the expected returns and expected dividend growth are stationary, deviations of the dividend–price ratio (dp_t) from its long-term mean ($\overline{dp}$) ought to forecast either future returns, or future dividend growth rates, or both.

This accounting identity has motivated some of the earliest empirical work in return predictability, which regressed returns on the lagged dividend–price ratio, as in Eq. (4):

$$(r_{t+1} - \bar{r}) = \kappa_r (dp_t - \overline{dp}) + \tau^r_{t+1} \, , \tag{4}$$

$$(\Delta d_{t+1} - \bar{d}) = \kappa_d (dp_t - \overline{dp}) + \tau^d_{t+1} \, , \tag{5}$$

$$(dp_{t+1} - \overline{dp}) = \phi (dp_t - \overline{dp}) + \tau^{dp}_{t+1} \, , \tag{6}$$

where $\bar{r}$ is the long-run mean return and τ^r is a mean-zero innovation. The logic of (3) suggests that the dividend–price ratio could predict future dividend growth rates instead of, or in addition to, future returns. Testing for dividend growth predictability would lead one to estimate Eq. (5), where $\bar{d}$ denotes the long-run mean log dividend growth.

The empirical return predictability literature started out by estimating Eq. (4) with the dividend–price ratio on the right-hand side; see [12,17,24,29,34,53] and [42], among others. It found evidence for return predictability, i. e., $\kappa_r > 0$. This finding was initially interpreted as evidence against the efficient market hypothesis.

Around the same time, [25] and [52] document a negative autocorrelation in long-horizon returns. Good past returns forecast bad future returns. [16] and [18] summarize the evidence based on long-horizon autocorrelations and variance ratios, and conclude that the statistical evidence in favor of mean reversion in long-horizon returns is weak, possibly due to small sample problems. This motivates [4] to use a large cross-section of countries and use a panel approach instead. They in turn document strong evidence in favor of mean-reversion of long-horizon returns with an estimated half-life of 3–3.5 years.

Second, other financial ratios, such as the earnings-price ratio or the book-to-market ratio, or macro-economic variables such as the consumption-wealth ratio, the labor income-to-consumption ratio, or the housing collateral ratio, as well as corporate decisions, and the cross-sectional price of risk have subsequently been shown to predict returns as well; see [3,38,39,43,45,50] and [51], among others.

Third, long-horizon returns are typically found to be more predictable than one-period ahead returns. The coefficient $\kappa_r(H)$ in the H-period regression

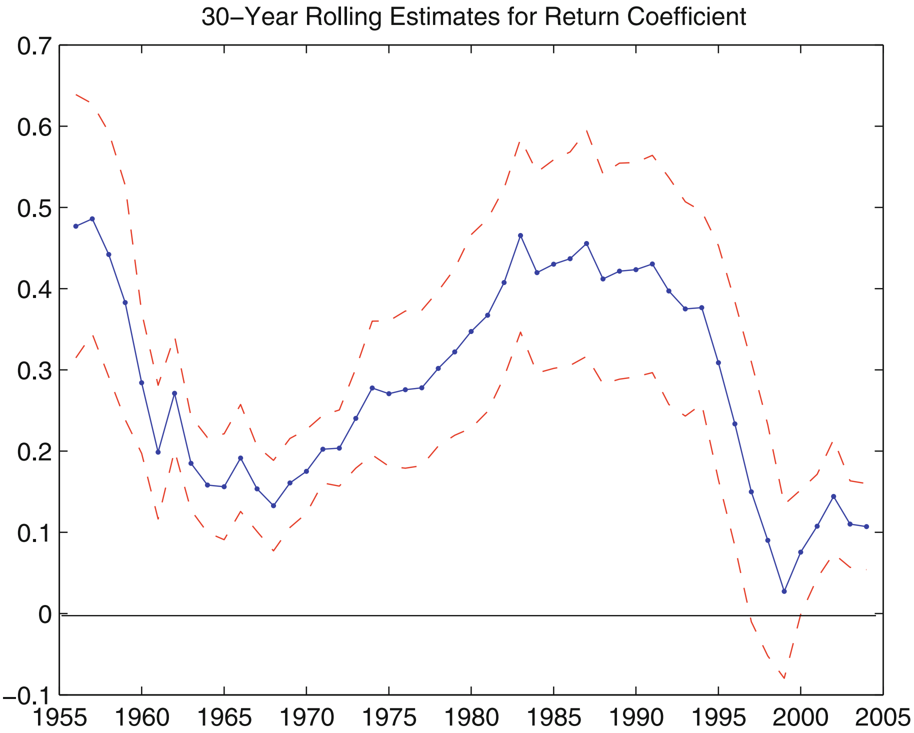
$$\sum_{j=1}^{H} r_{t+j} = \kappa_r(H) \, dp_t + \tau^r_{t,t+H} \tag{7}$$

exceeds the coefficient κ_r in the one-period regression. This finding is interpreted as evidence for the fact that the time-varying component in expected returns is quite persistent.

Fourth, these studies conclude that growth rates of fundamentals, such as dividends or earnings, are much less forecastable than returns using financial ratios. This suggests that most of the variation of financial ratios is due to variation in expected returns.

Fifth, predictability of stock returns does not only arise for the US. Studies by [10,26,33], and [2] analyze a large cross-section of countries and find evidence in favor of predictability by financial ratios in some countries, even though the evidence is mixed. More robust results are documented for the predictive ability of term structure variables.

These conclusions regarding predictability of stock returns are controversial because the forecasting relationship of financial ratios and future stock returns exhibits three disconcerting statistical features. First, correct inference is problematic because financial ratios are extremely persistent. The empirical literature typically augments Eq. (4) with an auto-regressive specification for the predictor variable, as in Eq. (6), where $\overline{dp}$ is the long-run mean of the dividend–price ratio. The estimated autoregressive



Financial Economics, Return Predictability and Market Efficiency, Figure 1
Parameter Instability in Return Predictability Coefficient

parameter ϕ is near unity and standard tests leave the possibility of a unit root open (i. e., $\phi = 1$). [2,27,46,55] and [58] conclude that the statistical evidence of forecastability is weaker once tests are adjusted for high persistence. [1,2,15,42,57] and [20] derive asymptotic distributions for predictability coefficients under the assumption that the forecasting variable follows a local-to-unit root, yet stationary, process.

Second, financial ratios have poor out-of-sample forecasting power, as shown in [7,31], and [32], but see [35] and [14] for different interpretations of the out-of-sample tests and evidence.

Third, the forecasting relationship of returns and financial ratios exhibits significant instability over time. Figure 1 shows that in rolling 30-year regressions of annual log CRSP value-weighted returns on lagged log dividend–price ratios, the ordinary least squares (OLS) regression coefficient varies between zero and 0.5 and the associated R^2 ranges from close to zero to 30% depending on the subsample.

The figure plots estimation results for the equation $r_{t+1} - \bar{r} = \kappa_r(dp_t - \overline{dp}) + \tau^r_{t+1}$. It shows the estimates for κ_r using 30-year rolling windows. The dashed line in the left panels denote the point estimate plus or minus one standard deviation. The standard errors are asymptotic. The parameters $\bar{r}$ and $\overline{dp}$ are the sample means of log returns r and the log dividend–price ratio dp. The data are annual for 1927–2004.

[60] and [49] report evidence in favor of breaks in the OLS coefficient in the forecasting regression of returns on the lagged dividend–price ratio, while [41] report evidence for structural shifts in $\overline{dp}$. [47] use Bayesian methods to estimate structural breaks in the equity premium.

Empirical Evidence Revisited

Table 1 reviews the empirical evidence using annual value-weighted CRSP log return, dividend growth, and dividend–price ratio data for 1927–2004. In Panel A, the system of Eqs. (4) and (5) is estimated by GMM. The first row indicates that a higher dividend–price ratio leads to a higher return ($\kappa_r = .094$ in Column 2) and a higher dividend growth rate ($\kappa_d = .005$ in Column 1). The latter coefficient has the wrong sign, but the coefficient is statistically indistinguishable from zero. The asymptotic standard error on the estimate for κ_r is .046. The corresponding asymptotic p-value is 3.6% so that κ_r is statistically different from zero at conventional levels. In other words, the

Financial Economics, Return Predictability and Market Efficiency, Table 1
Return and Dividend Growth Predictability in the Data

	κ_d	κ_r	ϕ	PV violation
Panel A: No Long-Horizon Moments $H = \{1\}$				
No Break	.005	.094	.945	−.046
	(.037)	(.046)	(.052)	
1 Break ('91)	.019	.235	.813	.004
	(.047)	(.055)	(.052)	
2 Breaks ('54, '94)	.124	.455	.694	−.001
	(.073)	(.079)	(.070)	
Panel B: Long-Horizon Moments $H = \{1, 3, 5\}$				
No Break	.021	.068	.990	.189
	(.018)	(.038)	(.032)	
1 Break ('91)	.012	.210	.834	.076
	(.019)	(.043)	(.042)	
2 Breaks ('54, '94)	.080	.409	.697	.100
	(.065)	(.078)	(.060)	

dividend–price ratio seems to predict stock returns, but not dividend growth. A similar result holds if returns in excess of a risk-free rate are used, or real returns instead of nominal returns.

[41] conduct an extensive Monte Carlo analysis to investigate the small-sample properties of estimates for κ_r and κ_d. Consistent with [55], the estimate for κ_r displays an upward small-sample bias. In addition, the standard error on κ_r is understated by the asymptotic standard error. As a result, one can no longer reject the null hypothesis that κ_r is zero. Based on this evidence, one is tempted to conclude that neither returns nor dividend growth are forecastable.

The second and third rows implement the suggestion of [41] to correct the long-run mean dividend–price ratio, $\overline{dp}$, for structural breaks. The data strongly suggest either one break in 1991, or two breaks in 1954 and 1994 in favor of either no breaks or three breaks. This break-adjusted dividend–price ratio is less persistent and less volatile. Its lower persistence alleviates the econometric issues mentioned above.

The second row of Table 1 uses the one-break adjusted dividend–price ratio as a regressor in the return and dividend growth predictability equations. The evidence in favor of return predictability is substantially strengthened. The point estimate for κ_r more than doubles to .235, and is highly significant. In the two-break case in the third row, the point estimate further doubles to 0.455. The small-sample bias in κ_r is negligible relative to the size of the coefficient. The R^2 of the return equation is 10% in the one-break case and even 23% in the two-break case. This compares to 3.8% in the no-break case. Furthermore, rolling regression estimates of κ_r indicate that it is much more stable over time when the break-adjusted dp series is used as a regressor. The dividend growth coefficient κ_d remains statistically indistinguishable from zero. This evidence strengthens the view that returns are predictable and dividend growth is not, and that these findings are not an artefact of statistical issues.

This table reports GMM estimates for the parameters $(\kappa_d, \kappa_r, \phi)$ and their asymptotic standard errors (in parentheses). The results in panel A are for the system with one-year ahead equations for dividend growth and returns ($H = 1$, $N = 0$). The results in panel B are for the system with one-year, three-year and five-year ahead equations for dividend growth and returns ($H = \{1, 3, 5\}$, $N = 2$). The first-stage GMM weighting matrix is the identity matrix. The asymptotic standard errors and p-values are computed using the Newey–West HAC procedure (second stage weighting matrix) with four lags in panel A and $H = 5$ lags in panel B. The last column denotes the present-value constraint violation of the univariate OLS slope estimators: $(1-\rho\phi^{\text{ols}})^{-1}(\kappa_r^{\text{ols}}-\kappa_d^{\text{ols}})$. It is expressed in the same units as κ_d and κ_r. In panel B this number is the average violation of the three constraints, one constraint at each horizon. The dividend–price ratio in rows 1 and 4 is the unadjusted one. In rows 2 and 5, the dividend–price ratio is adjusted for one break in 1991, and in rows 3 and 6, it is the series adjusted for two breaks in 1954 and 1994. All estimation results are for the annual sample 1927–2004.

Structural Model

What are researchers estimating when they run the return predictability regression (4)? How are the return and dividend growth predictability regressions in (4) and (5) related? To answer these important questions, we set up a simple structural model with time-varying expected returns and expected dividend growth rates. This structural model has the system of Eqs. (4)–(6) as its reduced-form. The main purpose of this model is to show that (i) the dividend–price ratio is a contaminated predictor of returns and dividend growth rates, (ii) that the parameters in (4)–(6) have to satisfy a cross-equation restriction, which we call the *present-value constraint*, and (iii) this restriction enables decomposing the dividend–price ratio into expected returns and expected dividend growth. Similar models can be derived for financial ratios other than the dividend–price ratio (e. g., [61]). [6] show how stock returns and book-to-market ratios are related in a general equilibrium model.

A Present-Value Model

We assume that expected dividend growth, z, and expected returns, x, follow an AR(1) process with autoregressive coefficient ϕ:

$$\Delta d_{t+1} - \bar{d} = z_t + \epsilon_{t+1}\,, \qquad z_{t+1} = \phi z_t + \zeta_{t+1}\,, \quad (8)$$

$$r_{t+1} - \bar{r} = x_t + \eta_{t+1}\,, \qquad x_{t+1} = \phi x_t + \xi_{t+1}\,. \quad (9)$$

The model has three fundamental shocks: an innovation in unexpected dividends ϵ_{t+1}, an innovation in expected dividends ζ_{t+1}, and an innovation in expected returns ξ_{t+1}. We assume that all three errors are serially uncorrelated and have zero cross-covariance at all leads and lags: $\text{Cov}(\epsilon_{t+1}, \zeta_{t+j}) = 0,\ \forall j \neq 1, \text{Cov}(\xi_{t+1}, \zeta_{t+j}) = 0,\ \forall j \neq 1$, and $\text{Cov}(\epsilon_{t+1}, \xi_{t+j}) = 0,\ \forall j$, except for a contemporaneous correlation between expected return and expected dividend growth innovations $\text{Cov}(\zeta_t, \xi_t) = \chi$, and a correlation between expected and unexpected dividend growth innovations $\text{Cov}(\zeta_t, \epsilon_t) = \lambda$. We discuss innovations to unexpected returns η below.

In steady-state, the log dividend–price ratio is a function of the long-run mean return and dividend growth rate $\overline{dp} = \log\left((\bar{r} - \bar{d})/(1 + \bar{d})\right)$. The log dividend–price ratio in (3) can then be written as:

$$dp_t - \overline{dp} = \frac{x_t - z_t}{1 - \rho\phi}\,. \quad (10)$$

The dividend–price ratio is the difference of two AR(1) processes with the same root ϕ, which is again an AR(1) process. I.e., we recover Eq. (6).

The return decomposition in [9] implies that the innovation to unexpected returns follows from the three fundamental shocks (i. e., combine (2) with (8)–(10)):

$$\eta_{t+1} = \frac{-\rho}{1 - \rho\phi}\xi_{t+1} + \frac{\rho}{1 - \rho\phi}\zeta_{t+1} + \epsilon_{t+1}\,. \quad (11)$$

Since both ρ and ϕ are positive and $\rho\phi < 1$, a positive shock to expected returns leads, ceteris paribus, to a negative contemporaneous return. Likewise, a shock to expected or unexpected dividend growth induces a positive contemporaneous return.

Contaminated Predictor

The first main insight from the structural model is that the demeaned dividend–price ratio in (10) is an imperfect forecaster of both returns and dividend growth. Returns are predicted by x_t (see Eq. (9)), but variation in the dividend–price ratio is not only due to variation in x, but also in expected dividend growth z_t. The same argument applies to dividend growth which is predicted by z_t (see Eq. (8)). This implies that the regressions in the reduced-form model in (4) and (5) suffer from an errors-in-variables problem [24,30,37].

To illustrate the bias, we can link the regression coefficients κ_r and κ_d explicitly to the underlying structural parameters:

$$\kappa_r = \frac{\text{Cov}(r_{t+1}, dp_t)}{\text{Var}(dp_t)} = \frac{(1 - \rho\phi)(\sigma_\xi^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,, \quad (12)$$

$$\kappa_d = \frac{\text{Cov}(\Delta d_{t+1}, dp_t)}{\text{Var}(dp_t)} = \frac{-(1 - \rho\phi)(\sigma_\zeta^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,. \quad (13)$$

If growth rates are constant, i. e., $\chi = 0$ and $\sigma_\zeta = 0$, then the dividend–price ratio is a perfect predictor of returns and $\kappa_r^\star = 1 - \rho\phi$. In all other cases, there is a bias in the return predictability coefficient:

$$\kappa_r^\star - \kappa_r = \frac{(1 - \rho\phi)(\sigma_\zeta^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,. \quad (14)$$

[24] argue that κ_r is downward biased ($\kappa_r^\star - \kappa_r > 0$). In fact, the structural parameters that are implied by the reduced-form model parameters indicate an upward bias. This occurs because the correlation between expected dividend growth and expected returns is sufficiently high.

A similar argument applies to κ_d. [40] construct a variable based on the co-integrating relationship between consumption, dividends from asset wealth, and dividends from human wealth. They show that this variable has strong predictive power for dividend growth, and they show that expected returns and expected growth rates are highly positively correlated. This implies that expected growth rates and expected returns have an offsetting effect on financial ratios, which makes it hard to reliably detect time-varying growth rates using such financial ratios.

Present-Value Constraint

The second main insight from the structural model is that there is a cross-equation restriction on the three innovations $\tau = (\tau^d, \tau^r, \tau^{dp})$ of the reduced-form model (4)–(6). Expressed in terms of the structural parameters, these innovations are:

$$\tau_{t+1}^d = \epsilon_{t+1} + x_t\left(\frac{-\kappa_d}{1 - \rho\phi}\right) + z_t\left(\frac{\kappa_r}{1 - \rho\phi}\right) \quad (15)$$

$$\tau_{t+1}^r = \epsilon_{t+1} + x_t\left(\frac{-\kappa_d}{1 - \rho\phi}\right) + z_t\left(\frac{\kappa_r}{1 - \rho\phi}\right) - \rho\left(\frac{\xi_{t+1} - \zeta_{t+1}}{1 - \rho\phi}\right) \quad (16)$$

$$\tau_{t+1}^{dp} = \frac{\xi_{t+1} - \zeta_{t+1}}{1 - \rho\phi} \,. \tag{17}$$

They imply the present value restriction:

$$\rho\tau_{t+1}^{dp} = \tau_{t+1}^{d} - \tau_{t+1}^{r} \;\Leftrightarrow\; \kappa_r - \kappa_d = 1 - \rho\phi \,. \tag{18}$$

Another way to write this restriction is as a restriction on a weighted sum of κ_r and κ_d: Any two equations from the system (4)–(6) implies the third. Evidence that dividend growth is not forecastable is evidence that returns are forecastable: if $\kappa_d = 0$ in Eq. (18), then $\kappa_r > 0$ because $\rho\phi < 1$. If estimating (5) uncovers that a high dividend–price ratio forecasts a higher future dividend growth ($\kappa_d > 0$), as we showed it does, then this strengthens the evidence for return predictability. [19] makes an important and closely related point: That it is important to impose the present-value relationship when testing the null hypothesis of no return predictability. That null ($\kappa_r = 0$) is truly a joint hypothesis, because it implies a negative coefficient in the dividend growth equation ($\kappa_d < 0$). [19], too, finds strong evidence for return predictability.

Returning to Panel A of Table 1, Column 3 backs out the AR(1) coefficient ϕ from the estimated κ_d and κ_r, and from the present-value constraint (18).[1] In the first row, $\phi = .945$, and is statistically undistinguishable from a unit root. This high persistence is a familiar result in the literature. The last column reports the left-hand side and the right-hand side of Eq. (18) for *univariate* OLS regressions of (4)–(6). It shows the violation of the present-value constraint. In the first row, the violation is half as large as the actual point estimate κ_r. The standard OLS point estimates do not satisfy the present-value constraint, which can lead to faulty inference.

However, when we use the break-adjusted dividend–price ratio series in rows 2 and 3, we find that (1) the persistence of the break-adjusted dp ratio is much lower than the unadjusted series (.81 and .69 versus .95), and (2) the present-value constraint is satisfied by the OLS coefficients.

A similar present-value constraint can be derived for long-horizon return and dividend growth regressions:

$$\kappa_r(H) = \kappa_r\left(\frac{1-\phi^H}{1-\phi}\right)$$
$$\kappa_d(H) = \kappa_d\left(\frac{1-\phi^H}{1-\phi}\right) .$$

Not only are the coefficients on the long-horizon return predictability regressions for all horizons linked to each other (see [8]), all long-horizon regression coefficients in the return equations are also linked to those from the dividend growth equations. I. e., there is one present-value constraint for each horizon H. Imposing these restrictions in a joint estimation procedure improves efficiency.

Panel B of Table 1 shows the results from a joint estimation of 1-year, 3-year, and 5-year cumulative returns and dividend growth rates on the lagged dividend–price ratio. Because of the restrictions, there are only two parameters to be estimated from these six equations. The results are close to those from the one-year system in Panel A, confirming the main message of [8]. The main conclusion remains that returns are strongly predictable, and dividend growth rates are not.

Exploiting Correlation in Innovations

The present-value model implies a restriction on the innovations in returns and the dividend–price ratio (see Eq. (18)). A third main insight from the structural model is that this correlation contains useful information for estimating the structural parameters, and hence for how much return predictability and dividend growth predictability there truly is. [48] show that exploiting the correlation between expected and unexpected stock returns can lead to substantially more accurate estimates. The information in correlations is incorporated by specifying a prior belief about the correlation between expected and unexpected returns, and updating that prior in a Bayesian fashion using observed data. Their method ignores the present-value constraint. The structural parameters in Panel B of Table 1, which impose the present-value constraint, imply that two-thirds of the variability in the price-dividend ratio is due to expected future returns and one-third is due to expected future dividend growth rates.

Likewise, [59] write down a model like (8)–(9) where expected returns and growth rates of dividends are autoregressive, exploiting the present-value constraint. Because the price-dividend ratio is linear in expected returns x and expected dividend growth z (see Eq. (10)), its innovations in (17) can be attributed to either innovations in expected returns or expected growth rates. The present-value constraint enables one to disentangle the information in price-dividend ratios about both expected returns and growth rates, and therefore to undo the contamination coming from correlated innovations. With this decomposition in hand, it is then possible to recover the full time-series of expected returns, x, and expected growth rates, z. [59] show that the resulting processes are strong predictors of realized returns and realized dividend growth rates, respectively. This underscores

[1] The linearization parameter ρ is tied to the average dividend–price ratio, and is held fixed at 0.9635.

the importance of specifying a present-value model to address return predictability.

Geometric or Arithmetic Returns

As a final comment, most predictive regressions are estimated using geometric, i. e. log returns, instead of arithmetic, i. e. simple returns. This choice is predominantly motivated by the [12] log-linearization discussed before. Since investors are ultimately interested in arithmetic instead of log returns, [59] specify a process for expected simple returns instead. This is made possible by applying the techniques of linearity-inducing models, recently introduced by [28].

Future Directions

The efficient market hypothesis, which states that markets efficiently aggregate all information, was first interpreted to mean that returns are not predictable. Early evidence of predictability of stock returns by the lagged dividend–price ratio seemed to be evidence against the efficient market hypothesis. However, return predictability and efficient markets are not incompatible because return predictability arises naturally in a world with time-varying expected returns. In the last 15 years, the empirical literature has raised a set of statistical objections to return predictability findings. Meanwhile, the theoretical literature has progressed, seemingly independently, in its pursuit of new ways to build models with time-varying expected returns. Only very recently has it become clear that theory is necessary to understand the empirical facts.

In this entry, we have set up a simple present-value model with time-varying expected returns that generates the regression that is the focus of the empirical literature. The model also features time-varying expected dividend growth. It shows that the dividend–price ratio contains information about both expected returns and expected dividend growth. A regression of returns on the dividend–price ratio may therefore be a poor indicator of the true extent of return predictability. At the same time, the present-value model provides a solution to this problem: It disentangles the two pieces of information in the price-dividend ratio. This allows us to interpret the standard predictability regressions in a meaningful way. Combining data with the present-value model, we conclude that there is strong evidence for return predictability. We interpret this as evidence for the presence of time-varying expected returns, not evidence against the efficient market hypothesis. The main challenge for the future is to better understand the underlying reasons for this time-variation.

Bibliography

Primary Literature

1. Amihud Y, Hurvich CM (2004) Predictive regressions: A reduced-bias estimation method. Financial Quant Anal 39:813–841
2. Ang A, Bekaert G (2007) Stock return predictability: Is it there? Rev Financial Stud 20(3):651–707
3. Baker M, Wurgler J (2000) The equity share in new issues and aggregate stock returns. J Finance 55:2219–2258
4. Balvers R, Wu Y, Gilliland E (2000) Mean reversion across national stock markets and parametric contrarian investment strategies. J Finance 55:745–772
5. Bansal R, Yaron A (2004) Risks for the long-run: A potential resolution of asset pricing puzzles. J Finance 59(4):1481–1509
6. Berk JB, Green RC, Naik V (1999) Optimal investment, growth options and security returns. J Finance 54:1153–1607
7. Bossaerts P, Hillion P (1999) Implementing statistical criteria to select return forecasting models: What do we learn? Rev Financial Stud 12:405–428
8. Boudoukh J, Richardson M, Whitelaw RF (2007) The myth of long-horizon predictability. Rev Financial Stud (forthcoming)
9. Campbell JY (1991) A variance decomposition for stock returns. Econ J 101:157–179
10. Campbell JY (2003) Consumption-based asset pricing. In: Constantinides G, Harris M, Stulz R (eds) Handbook of the Economics of Finance. North-Holland, Amsterdam (forthcoming)
11. Campbell JY, Cochrane JH (1999) By force of habit: A consumption-based explanation of aggregate stock market behavior. J Political Econ 107:205–251
12. Campbell JY, Shiller RJ (1988) The dividend–price ratio and expectations of future dividends and discount factors. Rev Financial Stud 1:195–227
13. Campbell JY, Shiller RJ (1991) Yield spreads and interest rates: A bird's eye view. Rev Econ Stud 58:495–514
14. Campbell JY, Thompson S (2007) Predicting excess stock returns out of sample: Can anything beat the historical average? Rev Financial Stud (forthcoming)
15. Campbell JY, Yogo M (2002) Efficient tests of stock return predictability. Harvard University (unpublished paper)
16. Campbell JY, Lo AW, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
17. Cochrane JH (1991) Explaining the variance of price-dividend ratios. Rev Financial Stud 5(2):243–280
18. Cochrane JH (2001) Asset Pricing. Princeton University Press, Princeton
19. Cochrane JH (2006) The dog that did not bark: A defense of return predictability. University of Chicago Graduate School of Business (unpublished paper)
20. Eliasz P (2005) Optimal median unbiased estimation of coefficients on highly persistent regressors. Department of Economics, Princeton University (unpublished paper)
21. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–101
22. Fama EF (1970) Efficient capital markets: A review of theory and empirical work. J Finance 25:383–417
23. Fama EF (1991) Efficient markets: II. J Finance 46(5):1575–1618
24. Fama EF, French KR (1988) Dividend yields and expected stock returns. J Financial Econ 22:3–27

25. Fama EF, French KR (1988) Permanent and temporary components of stock prices. J Political Econ 96(2):246–273
26. Ferson WE, Harvey CR (1993) The risk and predictability of international equity returns. Rev Financial Stud 6:527–566
27. Ferson WE, Sarkissian S, Simin TT (2003) Spurious regressions in financial economics? J Finance 58(4):1393–1413
28. Gabaix X (2007) Linearity-generating processes: A modelling tool yielding closed forms for asset prices. MIT (working paper)
29. Goetzman WN, Jorion P (1993) Testing the predictive power of dividend yields. J Finance 48:663–679
30. Goetzman WN, Jorion P (1995) A longer look at dividend yields. J Bus 68:483–508
31. Goyal A, Welch I (2003) Predicting the equity premium with dividend ratios. Manag Sci 49(5):639–654
32. Goyal A, Welch I (2006) A comprehensive look at the empirical performance of the equity premium prediction. Rev Financial Stud (forthcoming)
33. Hjalmarsson E (2004) On the predictability of global stock returns, Yale University (unpublished paper)
34. Hodrick R (1992) Dividend yields and expected stock returns: Alternative procedures for inference and measurement. Rev Financial Stud 5:357–386
35. Inoue A, Kilian L (2004) In-sample or out-of-sample tests of predictability: Which one should we use? Econom Rev 23: 371–402
36. Jensen MC (1978) Some anomalous evidence regarding market efficiency. J Financial Econ 6:95–101
37. Kothari S, Shanken J (1992) Stock return variation and expected dividends: A time-series and cross-sectional analysis. J Financial Econ 31:177–210
38. Lamont O (1998) Earnings and expected returns. J Finance 53:1563–87
39. Lettau M, Ludvigson SC (2001) Consumption, aggregate wealth and expected stock returns. J Finance 56(3):815–849
40. Lettau M, Ludvigson SC (2005) Expected returns and expected dividend growth. J Financial Econ 76:583–626
41. Lettau M, Van Nieuwerburgh S (2006) Reconciling the return predictability evidence. Rev Financial Stud (forthcoming)
42. Lewellen JW (2004) Predicting returns with financial ratios. J Financial Econ 74(2):209–235
43. Lustig H, Van Nieuwerburgh S (2005) Housing collateral, consumption insurance and risk premia: An empirical perspective. J Finance 60(3):1167–1219
44. Lustig H, Van Nieuwerburgh S (2006) Can housing collateral explain long-run swings in asset returns? University of California at Los Angeles and New York University (unpublished manuscript)
45. Menzly L, Santos T, Veronesi P (2004) Understanding predictability. J Political Econ 112(1):1–47
46. Nelson CC, Kim MJ (1993) Predictable stock returns: The role of small sample bias. J Finance 43:641–661
47. Pastor L, Stambaugh RF (2001) The equity premium and structural breaks. J Finance 56(4):1207–1239
48. Pastor L, Stambaugh RF (2006) Predictive systems: Living with imperfect predictors, graduate School of Business. University of Chicago Journal of Finance (forthcoming)
49. Paye BS, Timmermann A (2006) Instability of return prediction models. J Empir Finance 13(3):274–315
50. Piazzesi M, Schneider M, Tuzel S (2007) Housing, consumption, and asset pricing. J Financial Econ 83(March):531–569
51. Polk C, Thompson S, Vuolteenaho T (2006) Cross-sectional forecasts of the equity risk premium. J Financial Econ 81: 101–141
52. Poterba JM, Summers LH (1988) Mean reversion in stock returns: Evidence and implications. J Financial Econ 22:27–60
53. Rozeff MS (1984) Dividend yields are equity risk premia. J Portfolio Manag 49:141–160
54. Shiller RJ (1984) Stock prices and social dynamics. Brook Pap Econ Act 2:457–498
55. Stambaugh RF (1999) Predictive regressions. J Financial Econ 54:375–421
56. Summers LH (1986) Does the stock market rationally reflect fundamental values? J Finance 41:591–601
57. Torous W, Volkanov R, Yan S (2004) On predicting returns with nearly integrated explanatory variables. J Bus 77:937–966
58. Valkanov R (2003) Long-horizon regressions: Theoretical results and applications. J Financial Econ 68:201–232
59. van Binsbergen J, Koijen RS (2007) Predictive regressions: A present-value approach. Duke University (working paper)
60. Viceira L (1996) Testing for structural change in the predictability of asset returns. Harvard University (unpublished manuscript)
61. Vuolteenaho T (2000) Understanding the aggregate book-market ratio and its implications to current equity-premium expectations. Harvard University (unpublished paper)

Books and Reviews

Campbell JY, Lo AW, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton

Cochrane JH (2005) Asset Pricing. Princeton University Press, Princeton, NJ

Malkiel BG (2004) A Random Walk Down Wall Street. W.W. Norton, New York

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model

RALITSA PETKOVA
Mays Business School, Texas A&M University,
College Station, USA

Article Outline

Glossary

Market capitalization Market capitalization is a measure of the size of a public company. It is equal to the share price times the number of shares outstanding. Small stocks have small market capitalizations, while large stocks have large market capitalizations.

Book-to-market ratio A ratio used to compare a company's book value to its market capitalization value. It is calculated by dividing the latest book value by the latest market value of the company.

Value stocks Value stocks tend to trade at lower prices relative to fundamentals like dividends, earnings, sales and others. These stocks are considered undervalued by value investors. Value stocks usually have high dividend yields, and high book-to-market ratios.

Growth stocks Growth stocks tend to trade at higher prices relative to fundamentals like dividends, earnings, sales and others. Growth stocks usually do not pay dividends and have low book-to-market ratios.

Market beta The market beta is a measure of the systematic risk of a security in comparison to the market as a whole. It measures the tendency of the security return to respond to market movements.

Capital asset pricing model (CAPM) The CAPM describes the relationship between risk and expected return and it is used in the pricing of risky securities. According to the CAPM, the expected return of a security equals the rate on a risk-free security plus a risk premium that increases in the security's market beta.

Definition of the Subject

Different stocks have different expected rates of return and many asset pricing models have been developed to understand why this is the case. According to such models, different assets earn different average returns because they differ in their exposures to systematic risk factors in the economy. Fama and French [12] derive a model in which the systematic risk factors are the market index, and two portfolios related to the size of a company, and its ratio of book value to market value (book-to-market). The size and book-to-market factors are empirically motivated by the observation that small stocks and stocks with high book-to-market ratios (value stocks) earn higher average returns than justified by their exposures to market risk (beta) alone. These observations suggest that size and book-to-market may be proxies for exposures to sources of systematic risk different from the market return.

Introduction

An important class of asset pricing models in finance are linear beta models. They assume that the expected return of an asset in excess of the risk-free rate is a linear function of exposures to systematic sources of risk. Usually, the asset's exposures to common sources of risk in the economy are referred to as betas. In general, linear beta models assume the following form for the unconditional expected excess return on assets:

$$E(R_i) = \gamma_M \beta_{i,M} + \sum \gamma_K \beta_{i,K}, \quad \text{for all} \quad i \tag{1}$$

where $E(R_i)$ is the expected excess return of asset i, γ_M is the market risk premium or the price for bearing market risk, and γ_K is the price of risk for factor K. The model stated above implies that exposures to systematic sources of risk are the only determinants of expected returns. Thus, assets with high betas earn higher expected returns. The betas are the slope coefficients from the following return-generating process:

$$R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + \sum \beta_{i,K} K_t + \varepsilon_{i,t}, \quad \text{for all} \quad i \tag{2}$$

where $R_{i,t}$ is the return on asset i in excess of the risk-free rate at the end of period t, $R_{M,t}$ is the excess return on the market portfolio at the end of period t, and K_t is the realization for factor K at the end of period t.

One approach of selecting the pervasive risk factors is based on empirical evidence. For example, many empirical studies document that small stocks have higher average returns than large stocks, and value stock have higher average returns than growth stocks (see [12] for a review). The differences in average returns of these classes of stocks are statistically and economically significant. If the market sensitivities of small and value stocks were high then their high average returns would be consistent with the Capital Asset Pricing Model (CAPM), which predicts that the market beta is the only determinant of average returns. However, the patterns in returns for these stocks cannot be explained by the CAPM.

In a series of papers, Fama and French [12,13,14] show that a three-factor model performs very well at capturing the size and value effects in average stock returns. The three factors are the excess return on the market portfolio, the return on a portfolio long in value stocks and short in growth stocks, and the return on a portfolio long in small stocks and short in large stocks.

The impressive performance of the Fama–French three-factor model has spurred an enthusiastic debate in the finance literature over what underlying economic interpretation to give to the size and book- to-market factors. One side of the debate favors a risk-based explanation and contends that these factors reflect systematic risks that the static CAPM has failed to capture. For example, if the return distributions of different assets change over time (i. e., expected returns, variances, correlation), then the investment opportunity set available to investors varies over time as well. If individual assets covary with variables that track this variation then the expected returns of these assets will reflect that. Fama and French argue that the factors in their model proxy for such variables.

Another side of the debate favors a non-risk explanation. For example, Lakonishok, Shleifer, and Vishny [22] argue that the book-to-market effect arises since investors over-extrapolate past earnings growth into the future and overvalue companies that have performed well in the past. Namely, investors tend to over-extrapolate recent performance: they overvalue the firms with good recent performance (growth) and undervalue the firms with bad recent performance (value). When the market realizes its mistake, the prices of the former fall, while the prices of the latter rise. Therefore on average, growth firms tend to underperform value firms. Daniel and Titman [9] suggest that stocks characteristics, rather than risks, are priced in the cross-section of average returns. Other authors attribute the success of the size and book-to-market factors to data-snooping and other biases in the data [21,27]. Berk, Green, and Naik [1] and Gomes, Kogan, and Zhang [17] derive models in which problems in the measurement of market beta may explain the Fama–French results.

This article focuses on the risk-based explanation behind the success of the Fama–French three-factor model. If the Fama–French factors are to be explained in the context of a rational asset pricing model, then they should be correlated with variables that characterize time variation in the investment opportunity set. The rest of the article is organized as follows. Section "The Fama–French Model as a Linear Beta Pricing Model" discusses the set-up of the Fama–French model and presents some empirical tests of the model. Section "Explaining the Performance of the Fama–French Model: A Risk-Based Interpretation" argues that the Fama–French factors proxy for fundamental variables that describe variation in the investment opportunity set over time, and presents empirical results. Section "Other Risk-Based Interpretations" presents additional arguments for the relation between the Fama–French factors and more fundamental sources of risk. Section "Future Directions" summarizes and concludes.

The Fama–French Model as a Linear Beta Pricing Model

Model Set-up

Fama and French [12] propose a three-factor linear beta model to explain the empirical performance of small and high book-to-market stocks. The intuition behind the factors they propose is the following.

If small firms earn higher average returns than large firms as a compensation for risk, then the return differential between a portfolios of small firms and a portfolio of large firms would mimic the factor related to size provided the two portfolios have similar exposures to other sources of risk. Similarly, if value firms earn higher average returns than growth firms as a compensation for risk, then the return differential between a portfolio of value firms and a portfolio of growth firms, would mimic the factor related to book-to-market provided the two portfolios have similar exposure to other sources of risk. Fama and French [12] construct two pervasive risk factors in this way that are now commonly used in empirical studies. The composition of these factors is explained below.

In June of each year independent sorts are used to allocate the NYSE, AMEX, and NASDAQ stocks to two size groups and three book-to-market groups. Big stocks are above the median market equity of NYSE firms and small stocks are below. Similarly, low book-to-market stocks are below the 30th percentile of book-to-market for NYSE firms, medium book-to-market stocks are in the middle 40 percent, and high book-to-market stocks are in the top 30 percent. Size is market capitalization at the end of June. Book-to-market is book equity at the last fiscal year end of the prior calendar year divided by market cap as of 6 months before formation. Firms with negative book equity are not considered. At the end of June of each year, six value-weight portfolios are formed, SL, SM, SH, BL, BM, and BH, as the intersections of the size and book-to-market groups. For example, SL is the value-weight return on the portfolio of stocks that are below the NYSE median in size and in the bottom 30 percent of book-to-market. The portfolios are rebalanced annually. *SMB* in each period is the difference between the equal-weight averages of the returns on the three small stock portfolios and the three big stock portfolios, constructed to be neutral with respect to book-to-market:

$$SMB = (SL + SM + SH)/3 - (BL + BM + BH)/3 \,. \quad (3)$$

Similarly, *HML* in each period is the difference between the return on a portfolio of high book-to-market stocks and the return on a portfolio of low book-to-market stocks, constructed to be neutral with respect to size:

$$HML = (SH + BH)/2 - (SL + BL)/2 \,. \quad (4)$$

Therefore, the Fama–French three-factor linear model implies that:

$$E(R_i) = \gamma_M \beta_{i,M} + \gamma_{SMB} \beta_{i,SMB} + \gamma_{HML} \beta_{i,HML} \,, \quad \text{for all} \quad i \quad (5)$$

where $E(R_i)$ is the excess return of asset *i*, γ_M is the market risk premium, γ_{SMB} is the price of risk for the size factor, and γ_{HML} is the price of risk for the book-to-market factor. The betas are the slope coefficients from the following return-generating process:

$$\begin{aligned} R_{i,t} &= \alpha_i + \beta_{i,M} R_{M,t} + \beta_{i,SMB} R_{SMB,t} \\ &\quad + \beta_{i,HML} R_{HML,t} + \varepsilon_{i,t} \,, \quad \text{for all} \quad i \end{aligned} \quad (6)$$

where $R_{i,t}$ is the return on asset *i* in excess of the risk-free rate at the end of period *t*, $R_{M,t}$ is the excess return on the market portfolio at the end of period *t*, $R_{SMB,t}$ is the return on the SMB portfolio at the end of period *t*, and $R_{HML,t}$ is the return on the *HML* portfolio at the end of period *t*.

Testing the Fama–French Model and Results

The return-generating process is Eq. (6) applies to the excess return of any asset. The Fama–French model is usually tested on a set of portfolios sorted by book-to-market and size. Similarly to the construction of *HML* and *SMB*, 25 value-weighted portfolios are formed as the intersections of five size and five book-to-market groups. These 25 portfolios are the test assets used most often in testing competing asset-pricing models. These assets represent one of the most challenging set of portfolios in the asset pricing literature.

In this article, monthly data for the period from July of 1963 to December of 2001 is used. The returns on the market portfolio, the risk-free rate, *HML*, and *SMB* are taken from Ken French's web site, as well as the returns on 25 portfolios sorted by size and book-to-market.

To test the Fama–French specification in Eq. (5), the Fama–MacBeth [15] cross-sectional method can be used. In the first pass of this method, a multiple time-series regression as in (6) is estimated for each one of the 25 portfolios mentioned above which provides estimates of the assets' betas with respect to the market return, and the size and book-to-market factors.

Table 1 reports the estimates of the factor loadings computed in the first-pass time-series regression (6) for each portfolio. The table also present joint tests of the significance of the corresponding loadings, computed from a seemingly unrelated regressions (SUR) system. This is done in order to show that the Fama–French factors are relevant in the sense that the 25 portfolios load significantly on them.

The results from Table 1 reveal that within each size quintile, the loadings of the portfolios with respect to *HML* increase monotonically with book-to-market. Within each size group, portfolios in the lowest book-to-market quintile (growth) have negative betas with respect to *HML*, while portfolios in the highest book-to-market quintile (value) have positive betas with respect to *HML*. Further, within each book-to-market quintile, the loadings of the portfolios with respect to *SMB* decrease monotonically with size. Within each book-to-market group, portfolios in the lowest size quintile (small) have positive betas with respect to *SMB*, while portfolios in the highest size quintile (large) have negative betas with respect to *SMB*. The table shows that small and large portfolios, and value and growth portfolios have similar market betas.

Note that only six of the 25 intercepts in Table 1 are significant (although the intercepts are jointly significant). The large R-square statistics show that the excess returns of the 25 portfolios are explained well by the three-factor model. Furthermore the large t-statistics on the size and book-to-market betas show that these factors contribute significantly to the explanatory power of the model.

The second step of the Fama–MacBeth procedure involves relating the average excess returns of the 25 portfolios to their exposures to the risk factors in the model. More specifically, the following cross-sectional relation is estimated

$$\begin{aligned} \overline{R}_{i,t} = \gamma_0 + \gamma_M \widehat{\beta}_{i,M} + (\gamma_{HML}) \widehat{\beta}_{i,HML} + (\gamma_{SMB}) \widehat{\beta}_{i,SMB} \\ + e_{i,t} \,. \end{aligned} \quad (7)$$

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 1
Loadings on the Fama–French Factors from Time-Series Regressions
This table reports loadings on the excess market return, R_M, and the Fama–French factors R_{HML} and R_{SMB} computed in time-series regressions for 25 portfolios sorted by size and book-to-market. The corresponding *t*-statistics are also reported and are corrected for autocorrelation and heteroscedasticity using the Newey–West estimator with five lags. The sample period is from July 1963 to December 2001. The intercepts are in percentage form. The last column reports *F*-statistics and their corresponding *p*-values from an SUR system, testing the joint significance of the corresponding loadings. The *p*-values are in percentage form. R^2s from each time-series regression are reported in percentage form

Regression: $R_{i,t} = \alpha_i + \beta_{i,M}R_{M,t} + \beta_{i,HML}R_{HML,t} + \beta_{i,SMB}R_{SMB,t} + \varepsilon_{i,t}$												
	Low	2	3	4	High		Low	2	3	4	High	
			α						t_α			F
Small	−0.38	0.01	0.04	0.18	0.12		−3.40	0.18	0.56	2.84	1.91	2.96
2	−0.17	−0.10	0.08	0.08	−0.00		−2.25	−1.45	1.15	1.28	−0.01	0.01
3	−0.07	−0.00	−0.09	0.01	0.00		−1.03	−0.03	−1.26	0.17	0.06	
4	0.16	0.21	−0.08	0.04	−0.05		1.67	−2.27	−0.99	0.61	−0.54	
Large	0.21	−0.04	−0.02	−0.09	−0.21		3.25	−0.53	−0.27	−1.29	−2.36	
			β_M						t_{β_M}			F
Small	1.04	0.96	0.93	0.92	0.98		44.38	39.40	50.88	46.60	43.39	> 100
2	1.11	1.03	1.00	0.99	1.08		48.84	45.42	46.47	60.69	52.11	< 0.01
3	1.09	1.07	1.03	1.01	1.10		52.59	38.53	32.93	52.70	38.97	
4	1.05	1.11	1.08	1.03	1.17		46.03	36.33	36.86	41.15	36.74	
Large	0.96	1.04	0.99	1.01	1.04		45.08	49.22	36.71	46.18	31.59	
			β_{HML}						$t_{\beta_{HML}}$			F
Small	−0.31	0.09	0.31	0.47	0.69		−5.86	1.79	9.62	14.97	17.10	> 100
2	−0.38	0.18	0.43	0.59	0.76		−8.52	2.96	7.36	13.97	23.28	< 0.01
3	−0.43	0.22	0.52	0.67	0.82		−14.90	3.10	7.39	10.58	15.94	
4	−0.45	0.26	0.51	0.61	0.83		−10.55	3.42	7.43	11.92	16.07	
Large	−0.38	0.14	0.27	0.64	0.85		−10.47	2.58	5.65	11.82	20.56	
			β_{SMB}						$t_{\beta_{SMB}}$			F
Small	1.41	1.33	1.12	1.04	1.09		36.39	24.68	36.50	24.34	25.40	> 100
2	1.00	0.89	0.75	0.70	0.82		27.61	18.51	15.90	25.31	25.68	< 0.01
3	0.72	0.51	0.44	0.38	0.53		24.97	7.68	6.81	8.28	8.87	
4	0.37	0.20	0.16	0.20	0.26		9.26	3.42	2.64	6.70	4.22	
Large	−0.26	−0.24	−0.24	−0.22	−0.08		−9.25	−6.92	−6.12	−6.81	−2.11	
						R^2						
				92.61	94.32	94.89	94.51	94.58				
				95.16	93.99	93.56	93.85	94.62				
				94.88	90.22	89.49	89.69	90.31				
				93.52	88.31	87.65	88.41	85.77				
				93.35	89.79	84.32	87.39	80.60				

The $\widehat{\beta}$ terms are the independent variables in the regression, while the average excess returns of the assets are the dependent variables. If loadings with respect to the Fama–French factors are important determinants of average returns, then there should be a significant price of risk associated with the factors.

Since the betas are estimated from the time-series regression in (6), they represent generated regressors in (7). This is the classical errors-in-variables problem, arising from the two-pass nature of this approach. Following Shanken [33], a correction procedure can be used that accounts for the errors-in-variables problem. Shanken's correction is designed to adjust for the overstated precision of the Fama–MacBeth standard errors. It assumes that the error terms from the time-series regression are independently and identically distributed over time, conditional on the time series of observations for the risk factors. The adjustment also assumes that the risk factors are generated by a stationary process. Jagannathan and Wang [19] argue that if the error terms are heteroscedastic, then the Fama–

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 2

Cross-Sectional Regressions with the Fama–French Factor Loadings

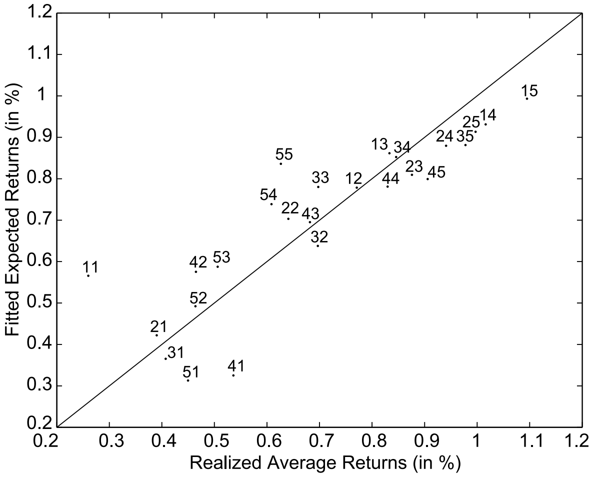
This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The sample period is from July 1963 to December 2001

The Fama–French Three-Factor Model					
	γ_0	γ_M	γ_{HML}	γ_{SMB}	Adj. R^2
Estimate	1.15	−0.65	0.44	0.16	71.00
FM *t*-stat	3.30	−1.60	3.09	1.04	
SH *t*-stat	3.19	−1.55	3.07	1.00	

MacBeth procedure does not necessarily result in smaller standard errors of the cross-sectional coefficients. In light of these two issues, researchers often report both unadjusted and adjusted cross-sectional statistics.

Table 2 reports the estimates of the factor prices of risk computed in the second-pass cross-sectional regression (7). The table also presents the t-statistics for the coefficients, adjusted for errors-in-variables following Shanken [33]. The table shows that the market beta is not an important factor in the cross-section of returns sorted by size and book-to-market.[1] Further, the table reveals that loadings on *HML* represent a significant factor in the cross-section of the 25 portfolios, even after correcting for the sampling error in the loadings. Loadings on *SMB* do not appear to be significant in the cross-section of portfolio returns for this time period. The large R-square of 0.71 shows that the loadings from the Fama–French model explain a significant portion of the cross-sectional variation in the average returns of these portfolios.

It is also helpful to examine the performance of the model visually. This is done by plotting the fitted expected return of each portfolio against its realized average return in Fig. 1. The fitted expected return is computed using the estimated parameter values from the Fama–French model specification. The realized average return is the time-series average of the portfolio return. If the fitted expected return and the realized average return for each portfolio are the same, then they should lie on a 45-degree line through the origin.



Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Figure 1

Fitted Expected Returns vs. Average Realized Returns for 1963:07-2001:12.

This figure shows realized average returns (%) on the *horizontal axis* and fitted expected returns (%) on the *vertical axis* for 25 size and book-to-market sorted portfolios. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile (1 being the smallest and 5 the largest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For each portfolio, the realized average return is the time-series average of the portfolio return and the fitted expected return is the fitted value for the expected return from the corresponding model. The straight line is the 45-degree line from the origin

Figure 1 shows the fitted versus realized returns for the 25 portfolios in two different models for the period from July of 1963 to December of 2001. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile of the portfolio (1 being the smallest and 5 the biggest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For example, portfolio 15 has the highest book-to-market value among the portfolios in the smallest size quintile. In other words, it is the smallest value portfolio.

It can be seen form the graph that the model goes a long way toward explaining the value effect: in general, the fitted expected returns on value portfolios (bigger second digit) are higher than the fitted expected returns on growth portfolios (lower second digit). This is consistent with the data on realized average returns for these portfolios. By inspection of Fig. 1, a few portfolios stand out as problematic for the FF model, in terms of distance from

[1] The estimate of the market risk premium tends to be negative. This result is consistent with previous results reported in the literature. Fama and French [11], Jagannathan and Wang [18], and Lettau and Ludvigson [24] report negative estimates for the market risk premium, using monthly or quarterly data.

the 45-degree line, namely the growth portfolios within the smallest and largest size quintiles (11, 41, and 51) and the value portfolios within the largest size quintiles (45, 54, and 55).

In summary, the Fama–French model performs remarkably well at explaining the average return difference between small and large, and value and growth portfolios.

The natural question that arises is what drives the superior performance of the Fama–French model in explaining average stock returns. One possible explanation is that the Fama–French factors *HML* and *SMB* proxy for sources of risk not captured by the return on the market portfolio. This explanation is consistent with a multifactor asset pricing model like the Intertemporal Capital Asset Pricing Model (ICAPM), which states that if investment opportunities change over time, then variables other than the market return will be important factors driving stock returns. Therefore, one possible interpretation of the *HML* and *SMB* portfolios is that they proxy for variables that describe how investment opportunities change over time. The following sections examine the ICAPM explanation behind the performance of the Fama–French model.

Explaining the Performance of the Fama–French Model: A Risk-Based Interpretation

The ICAPM Framework

The analysis in this paper assumes that asset returns are governed by the discrete-time version of the ICAPM of Merton [29]. According to the ICAPM, if investment opportunities change over time, then assets' exposures to these changes are important determinants of average returns in addition to the market beta. Campbell [3] develops a framework to model changes in the investment opportunity set as innovations in state variables that capture uncertainty about investment opportunities in the future. Therefore, the model for the unconditional expected excess returns on assets becomes

$$E(R_i) = \gamma_M \beta_{i,M} + \sum (\gamma_{u^K}) \beta_{i,u^K}, \quad \text{for all} \quad i \qquad (8)$$

where $E(R_i)$ is the excess return of asset i, γ_M is the market risk premium, and γ_{u^K} is the price of risk for innovations in state variable K. The betas are the slope coefficients from the following return-generating process:

$$R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + \sum (\beta_{i,u^K}) u_t^K + \varepsilon_{i,t}, \quad \text{for all} \quad i \quad (9)$$

where $R_{i,t}$ is the return on asset i in excess of the risk-free rate at the end of period t, $R_{M,t}$ is the excess return on the market portfolio at the end of period t, and u_t^K is the innovation to state variable K at the end of period t. The innovation is the unexpected component of the variable. According to the asset-pricing model, only the unexpected component of the state variable should command a risk premium. Note that the innovations to the state variables are contemporaneous to the excess market returns. This equation captures the idea that the market portfolio and the innovations to the state variables are the relevant risk factors.

It is important to specify a process for the time-series dynamics of the state variables in the model. A vector autoregressive (VAR) approach, for example, specifies the excess market return as the first element of a state vector z_t. The other elements of z_t are state variables that proxy for changes in the investment opportunity set. The assumption is that the demeaned vector z_t follows a first-order VAR:

$$z_t = \mathbf{A} z_{t-1} + \mathbf{u}_t \, . \qquad (10)$$

The residuals in the vector $\mathbf{u}_t$ are the innovations terms which are the risk factors in Eq. (2). These innovations are risk factors since they represent the surprise components of the state variables that proxy for changes in the investment opportunity set.

The State Variables of Interest

For the empirical implementation of the model described above, it is necessary to specify the identity of the state variables. Petkova [31] chooses a set of state variables to model the following aspects of the investment opportunity set: the yield curve and the conditional distribution of asset returns. In particular, she chooses the short-term Treasury bill, the term spread, the aggregate dividend yield, and the default spread.

The choice of these state variables is motivated as follows. The ICAPM dictates that the yield curve is an important part of the investment opportunity set. Furthermore, Long [28] points out that the yield curve is important in an economy with a bond market. Therefore, the short-term Treasury bill yield (*RF*) and the term spread (*TERM*) are good candidates that capture variations in the level and the slope of the yield curve. Litterman and Scheinkman [26] show that the two most important factors driving the term structure of interest rates are its level and its slope.

In addition to the yield curve, the conditional distribution of asset returns is a relevant part of the investment opportunity set facing investors in the ICAPM world. There is growing evidence that the conditional distribution of asset returns, as characterized by its mean and variance, changes over time. The time-series literature has identified variables that proxy for variation in the mean and variance of returns. The aggregate dividend yield (*DIV*), the default spread (*DEF*), and interest rates are among the most common.[2]

The variables described above are good candidates for state variable within the ICAPM. Merton [29] states that stochastic interest rates are important for changing investment opportunities. In addition, the default spread, the dividend yield, and interest rate variables have been used as proxies for time-varying risk premia under changing investment opportunities. Therefore, all these variables are likely to capture the hedging concerns of investors related the changes in interest rates and to variations in risk premia.

As argued in the previous sections of this article, two other variables proposed as candidates for state variables within the ICAPM are the returns on the *HML* and *SMB* portfolios. Fama and French [12] show that these factors capture common variation in portfolio returns that is independent of the market and that carries a different risk premium. The goal of the following section is to show that the FF factors proxy for the state variables described above that have been shown to track time-variation in the market risk premium and the yield curve.

Econometric Approach

First, a vector autoregressive (VAR) process for the vector of state variables is specified. The first element of the vector is the excess return on the market, while the other elements are *DIV*, *TERM*, *DEF*, *RF*, R_{HML}, and R_{SMB}, respectively. For convenience, all variables in the state vector have been demeaned. The first-order VAR is as follows:

$$\left\{ \begin{array}{c} R_{M,t} \\ DIV_t \\ TERM_t \\ DEF_t \\ RF_t \\ R_{HML,t} \\ R_{SMB,t} \end{array} \right\} = \mathbf{A} \left\{ \begin{array}{c} R_{M,t-1} \\ DIV_{t-1} \\ TERM_{t-1} \\ DEF_{t-1} \\ RF_{t-1} \\ R_{HML,t-1} \\ R_{SMB,t-1} \end{array} \right\} + \mathbf{u}_t \tag{11}$$

where $\mathbf{u}_t$ represents a vector of innovations for each element in the state vector. From $\mathbf{u}_t$ six surprise series can be extracted, corresponding to the dividend yield, the term spread, the default spread, the one-month T-bill yield, and the FF factors. They are denoted as follows: u^{DIV}, u^{TERM}, u^{DEF}, u^{RF}, u^{HML}, and u^{SMB}, respectively. This VAR represents a joint specification of the dynamics of all candidate state variables within the ICAPM. This specification treats the FF factors as potential candidates for state variables that command separate risk premia from the other variables.

The innovations derived from the VAR model are risk factors in addition to the excess return of the market portfolio. Asset's exposures to these risk factors are important determinants of average returns according to the ICAPM. To test the ICAPM specification, the Fama–MacBeth [15] cross-sectional method can be used as previously discussed. In the first pass of this method, a multiple time-series regression is specified which provides estimates of the assets' loadings with respect to the market return and the innovations in the state variables. More precisely, the following time-series regression is examined for each asset:

$$\begin{aligned} R_{i,t} = {} & \alpha_i + \beta_{i,M} R_{M,t} + (\beta_{i,\hat{u}^{DIV}})\hat{u}_t^{DIV} + (\beta_{i,\hat{u}^{TERM}})\hat{u}_t^{TERM} \\ & + (\beta_{i,\hat{u}^{DEF}})\hat{u}_t^{DEF} + (\beta_{i,\hat{u}^{RF}})\hat{u}_t^{RF} + (\beta_{i,\hat{u}^{HML}})\hat{u}_t^{HML} \\ & + (\beta_{i,\hat{u}^{SMB}})\hat{u}_t^{SMB} + \varepsilon_{i,t}, \quad \text{for all} \quad i\,. \end{aligned} \tag{12}$$

The $\hat{u}$-terms represent the estimated surprises in the state variables. Note that the innovations terms are generated regressors and they appear on the right-hand side of the equation. However, as pointed out by Pagan [30], the OLS estimates of the parameters' standard errors will still be correct if the generated regressor represents the unanticipated part of a certain variable. On the other hand, if the $\hat{u}$-terms are only noisy proxies for the true surprises in the state variables, then the estimates of the factor loadings in the above regression will be biased downwards. This will likely bias the results against finding a relation between the innovations and asset returns.

The second step of the Fama–MacBeth procedure involves relating the average excess returns of all assets to their exposures to the risk factors in the model. Therefore, the following cross-sectional relation applies

$$\begin{aligned} \overline{R}_{i,t} = {} & \gamma_0 + \gamma_M \widehat{\beta}_{i,M} + (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} + (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} \\ & + (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} + (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + (\gamma_{\hat{u}^{HML}})\widehat{\beta}_{i,\hat{u}^{HML}} \\ & + (\gamma_{\hat{u}^{SMB}})\widehat{\beta}_{i,\hat{u}^{SMB}} + e_{i,t}, \quad \text{for all} \quad t\,. \end{aligned} \tag{13}$$

Data, Time-Series Analysis, and Results

In this section, monthly data for the period from July of 1963 to December of 2001 is used. The state variables in the context of the ICAPM are the dividend yield of the value-weighted market index (computed as the sum of div-

[2] The following is only a partial list of papers that document time-variation in the excess market return and the variables they use: Campbell [2], term spread; Campbell and Shiller [4], dividend yield; Fama and Schwert [16], T-bill rate; Fama and French [10], default spread.

idends over the last 12 months, divided by the level of the index), the difference between the yield of a 10-year and a 1-year government bond (term spread), the difference between the yield of a long-term corporate Baa bond and a long-term government bond (default spread), and the one-month Treasury-bill yield. Data on bond yields is taken from the FRED® database of the Federal Reserve Bank of St. Louis. The T-bill yield and the term spread are used to measure the level and the slope of the yield curve, respectively.

VAR Estimation The state variables are the FF factors and the four predictive variables described above. All of them are included in a first-order VAR system. Campbell [3] emphasizes that it is hard to interpret estimation results for a VAR factor model unless the factors are orthogonalized and scaled in some way. In his paper the innovations to the state variables are orthogonal to the excess market return and to labor income. Following Campbell, the VAR system in Eq. (4) is triangularized in a similar way: the innovation in the excess market return is unaffected, the orthogonalized innovation in *DIV* is the component of the original *DIV* innovation orthogonal to the excess market return, and so on. The orthogonalized innovation to *DIV* is a change in the dividend/price ratio with no change in the market return, therefore it can be interpreted as a shock to the dividend. Similarly, shocks to the term spread, the default spread, the short-term rate, and the FF factors are orthogonal to the contemporaneous stock market return. As in Campbell [3], the innovations are scaled to have the same variance as the innovation in the excess market return.

It is interesting to note that the returns on the FF factors are very highly correlated with their respective innovation series. For example, the correlation between $R_{HML,t}$ and $\hat{u}_t^{HML}$ is 0.90, while the correlation between $R_{SMB,t}$ and $\hat{u}_t^{SMB}$ is 0.92. Therefore, the returns on the *HML* and *SMB* portfolios are good proxies for the innovations associated with those variables.

Relation Between R_{HML} and R_{SMB} and the VAR Innovations As a first step towards testing whether the FF factors proxy for innovations in state variables that track investment opportunities, the joint distribution of R_{HML} and R_{SMB} and innovations to *DIV*, *TERM*, *DEF*, and *RF* is examined. The following time-series regression is analyzed

$$\hat{u}_t = c_0 + c_1 R_{M,t} + c_2 R_{HML,t} + c_3 R_{SMB,t} + \varepsilon_t \qquad (14)$$

for each series of innovations in the state variables. The results for these regressions are presented in Table 3, with the corresponding t-statistics, below the coefficients, corrected for heteroscedasticity and autocorrelation. Innovations in the dividend yield, $\hat{u}_t^{DIV}$, covary negatively and significantly with the return on *HML*. In addition, $\hat{u}_t^{TERM}$ covaries positively and significantly with the *HML* return. These results are robust to the presence of the market factor in the regression. The return on the *HML* portfolio covaries positively and significantly with $\hat{u}_t^{DEF}$, while the return on the *SMB* factor covaries negatively with $\hat{u}_t^{DEF}$ (the corresponding t-statistic is marginally significant). The last regression in Table 3 indicates that the FF factors are not significant determinants of innovations in the T-bill yield. The results in the table remain unchanged if the independent variables in the equation above are the innovations to R_{HML} and R_{SMB} derived from the VAR system. The R-squares in the regressions reported in Table 3 are rather low. This does not imply, however, that the innovations in the state variables cannot potentially price assets as well as the FF factors. It could be the case that only the information in the FF factors correlated with the state variables is relevant for the pricing of risky assets. A similar point is made by Vassalou [34].

As pointed out by FF [10], the values of the term spread signal that expected market returns are low during expansions and high during recessions. In addition, FF document that the term spread very closely tracks the short-

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 3

Time-Series Regressions Showing the Contemporaneous Relations Between Innovations in State Variables and the Fama–French Factors

This table presents time-series regressions of innovations in the dividend yield ($\hat{u}_t^{DIV}$), term spread ($\hat{u}_t^{TERM}$), default spread ($\hat{u}_t^{DEF}$), and one-month T-bill yield ($\hat{u}_t^{RF}$) on the excess market return, R_M, and the Fama–French factors R_{HML} and R_{SMB}. The innovations to the state variables are computed in a VAR system. The t-statistics are below the coefficients and are corrected for heteroscedasticity and autocorrelation using the Newey–West estimator with five lags. The Adjusted R^2 is reported in percentage form. The sample period is from July 1963 to December 2001

Regression: $\hat{u}_t = c_0 + c_1 R_{M,t} + c_2 R_{HML,t} + c_3 R_{SMB,t} + \varepsilon_t$					
Dep. Variable	c_0	c_1	c_2	c_3	Adj. R^2
$\hat{u}_t^{DIV}$	0.00	−0.08	−0.30	−0.01	3.00
	0.85	−0.70	−2.43	−0.09	
$\hat{u}_t^{TERM}$	−0.00	0.06	0.24	0.03	2.00
	−0.56	0.75	2.30	0.59	
$\hat{u}_t^{DEF}$	−0.00	0.07	0.17	−0.12	2.00
	−0.38	1.11	2.10	−1.92	
$\hat{u}_t^{RF}$	0.00	−0.04	−0.13	0.01	0.00
	0.36	−0.51	−1.36	0.14	

term fluctuations in the business cycle. Therefore, positive shocks to the term premium are associated with bad times in terms of business conditions, while negative shocks are associated with good times. In light of the results documented by Petkova and Zhang [32], that value stocks are riskier than growth stocks in bad times and less risky during good times, the relation between *HML* and shocks to the term spread seems natural.

Another interpretation of the relation between shocks to the term spread and the *HML* portfolio is in the context of cash flow maturities of assets. This point is discussed by Cornell [8] and Campbell and Vuolteenaho [5]. The argument is that growth stocks are high-duration assets, which makes them similar to long-term bonds and more sensitive to innovations in the long end of the term structure. Similarly, value stocks have lower duration than growth stocks, which makes them similar to short-term bonds and more sensitive to shocks to the short end of the yield curve.

Chan and Chen [6] have argued that small firms examined in the literature tend to be marginal firms, that is, they generally have lost market value due to poor performance, they are likely to have high financial leverage and cash flow problems, and they are less likely to survive poor economic conditions. In light of this argument, it is reasonable to assume that small firms will be more sensitive to news about the state of the business cycle. Therefore, it is puzzling that I find no significant relation between *SMB* and surprises to the term spread. Innovations in the term spread seem to be mostly related to *HML*. This observation suggests that the *HML* portfolio might represent risk related to cash flow maturity, captured by unexpected movements in the slope of the term structure.

Innovations in default spread, u_t^{DEF}, stand for changes in forecasts about expected market returns and changes in forecasts about default spread. FF [10] show that the default premium tracks time variation in expected returns that tends to persist beyond the short-term fluctuations in the business cycle. A possible explanation for the negative relation between *SMB* and shocks to the default spread could be that bigger stocks are able to track long-run trends in the business cycle better than the smaller stocks. The result that *HML* is also related to shocks in the default spread is consistent with the interpretation of *HML* as a measure of distress risk. The distress risk interpretation of the book-to-market effect is advocated by FF [11,12,13,14] and Chen and Zhang [7], among others.

In summary, the empirical literature has documented that both value and small stocks tend to be under distress, with high leverage and cash flow uncertainty. The results in this study suggest that the book-to-market factor might be related to asset duration risk, measured by the slope of the term structure, while the size factor might be related to asset distress risk, measured by the default premium.

It is reasonable to test whether the significant relation between the state variables surprises and the FF factors gives rise to the significant explanatory power of *HML* and *SMB* in the cross-section of returns. The next section examines whether *HML* and *SMB* remain significant risk factors in the presence of innovations to the other state variables. The results from the cross-sectional regressions suggest that *HML* and *SMB* lose their explanatory power for the cross-section of returns once accounting for the other variables. This supports an ICAPM explanation behind the empirical success of the FF three-factor model.

Cross-Sectional Regressions

Incremental Explanatory Power of the Fama–French Factors This section examines the pricing performance of the full set of state variables considered before over the period from July 1963 to December 2001. The full set of state variables consists of the dividend yield, the term spread, the default spread, the short-term T-bill yield, and the FF factors. The innovations to these state variables derived from a VAR system are risk factors in the ICAPM model. The objective is to test whether an asset's loadings with respect to these risk factors are important determinants of its average return.

The first specification is

$$\begin{aligned}\overline{R}_{i,t} = \gamma_0 &+ \gamma_{MKT}\widehat{\beta}_{i,MKT} + (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} \\ &+ (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} + (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} \\ &+ (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + (\gamma_{\hat{u}^{HML}})\widehat{\beta}_{i,\hat{u}^{HML}} \\ &+ (\gamma_{\hat{u}^{SMB}})\widehat{\beta}_{i,\hat{u}^{SMB}} + e_{i,t}\,, \qquad (15)\end{aligned}$$

where the $\widehat{\beta}$ terms stand for exposures to the corresponding factor, while the γ terms stand for the reward for bearing the risk of that factor. The $\widehat{\beta}$ terms are the independent variables in the regression, while the average excess returns of the assets are the dependent variables. If loadings with respect to innovations in a state variable are important determinants of average returns, then there should be a significant price of risk associated with that state variable.

The results are reported in Table 4. The table shows that assets' exposures to innovations in R_{HML} and R_{SMB} are not significant variables in the cross-section in the presence of betas with respect to surprises in the other state variables. The corresponding *t*-statistics are 1.40 and 1.56, respectively, under the errors-in-variables correction. Therefore, based on the results presented in Table 4, the hypothesis that innovations in the dividend yield, the

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 4
Cross-Sectional Regressions Showing the Incremental Explanatory Power of the Fama–French Factor Loadings
This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The table presents results for the model including the excess market return, R_M, and innovations in the dividend yield, term spread, default spread, one-month T-bill yield, and the Fama–French factors *HML* and *SMB*. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The table examines the sample period from July 1963 to December 2001

The Model with Innovations in All State Variables									
	γ_0	γ_M	$\gamma_{\hat{u}DIV}$	$\gamma_{\hat{u}TERM}$	$\gamma_{\hat{u}DEF}$	$\gamma_{\hat{u}RF}$	$\gamma_{\hat{u}HML}$	$\gamma_{\hat{u}SMB}$	Adj. R^2
Estimate	1.11	–0.57	–0.83	3.87	0.37	–2.90	0.42	0.41	77.26
FM *t*-stat	3.29	–1.45	–0.94	3.53	0.42	–3.33	1.62	1.75	
SH *t*-stat	2.36	–1.10	–0.69	2.56	0.31	–2.44	1.40	1.56	

term spread, the default spread, and the short-term T-bill span the information contained in the FF factors cannot be rejected.

A Model Based on R_M, and Innovations in *DIV*, *TERM*, *DEF*, and *RF* This part examines separately the set of innovations in the variables associated with time-series predictability: the dividend yield, the term spread, the default spread, and the short-term T-bill. The model specification is as follows

$$\begin{aligned} R_{i,t} = \alpha_i &+ \beta_{i,M} R_{M,t} + (\beta_{i,\hat{u}^{DIV}})\hat{u}_t^{DIV} \\ &+ (\beta_{i,\hat{u}^{TERM}})\hat{u}_t^{TERM} + (\beta_{i,\hat{u}^{DEF}})\hat{u}_t^{DEF} \\ &+ (\beta_{i,\hat{u}^{RF}})\hat{u}_t^{RF} + \varepsilon_{i,t}, \quad \text{for all} \quad i \end{aligned} \tag{16}$$

$$\begin{aligned} \overline{R}_{i,t} = &\gamma_0 + \gamma_M \widehat{\beta}_{i,M} + (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} + (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} \\ &+ (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} + (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + e, \quad \text{for all} \quad t \end{aligned} \tag{17}$$

which corresponds to a model in which the relevant risk factors are innovations to predictive variables. The objective is to compare the pricing performance of this model with that of the Fama–French model for the cross-section of returns sorted by book-to-market and size. The specification is motivated by the previous observation that *HML* and *SMB* do not add explanatory power to the set of state variables that are associated with time-series predictability.

Table 5 report the estimates of the factor loadings computed in the first-pass time-series regressions defined in Eq. (16). It also presents joint tests of the significance of the corresponding loadings, computed from a SUR system. This is done in order to show that the innovations factors are relevant in the sense that the 25 portfolios load significantly on them. A similar analysis was performed on the Fama–French model in Sect. "The Fama–French Model as a Linear Beta Pricing Model".

An F-test implies that the 25 loadings on innovations to the term spread are jointly significant, with the corresponding p-value being 0.47%. Furthermore, portfolios' loadings on $\hat{u}_t^{TERM}$ are related to book-to-market: within each size quintile, the loadings increase monotonically from lower to higher book-to-market quintiles. In fact, the portfolios within the lowest book-to-market quintile have negative sensitivities with respect to $\hat{u}_t^{TERM}$, while the portfolios within the highest book-to-market quintile have positive loadings on $\hat{u}_t^{TERM}$. This pattern resembles very much the one observed in Table 1 for the loadings on R_{HML}.

Similarly, loadings on shocks to default spread are jointly significant in Table 5, with the corresponding p-value being 0.24%. Moreover, the slopes on $\hat{u}_t^{DEF}$ are systematically related to size. Within each book-to-market quintile, the loadings increase almost monotonically from negative values for the smaller size quintiles to positive values for the larger size quintiles. This pattern closely resembles the mirror image of the one observed in Table 1 for the loadings on R_{SMB}. The slopes on dividend yield and T-bill innovations do not exhibit any systematic patterns related to size or book-to-market. However, both of these are jointly significant.

Note that the R^2s in the time-series regressions with the innovations factors in Table 5 are smaller than the ones in the regressions with the FF factors in Table 1. This indicates that potential errors-in-variables problems that arise in measuring the factor loadings will be more serious in the case of the innovations terms. Therefore, the results will be potentially biased against finding significant factor loadings on the shocks to the predictive vari-

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 5

Loadings on R_M, $\hat{u}_t^{DIV}$, $\hat{u}_t^{TERM}$, $\hat{u}_t^{DEF}$, and $\hat{u}_t^{RF}$ from Time-Series Regressions

This table reports loadings on the excess market return, R_M, and innovations in the dividend yield ($\hat{u}_t^{DIV}$), term spread ($\hat{u}_t^{TERM}$), default spread ($\hat{u}_t^{DEF}$), and short-term T-bill ($\hat{u}_t^{RF}$) computed in time-series regressions for 25 portfolios sorted by size and book-to-market. The corresponding *t*-statistics are also reported and are corrected for autocorrelation and heteroscedasticity using the Newey–West estimator with five lags. The sample period is from July 1963 to December 2001. The last column reports *F*-statistics and their corresponding *p*-values from an SUR system, testing the joint significance of the corresponding loadings. The *p*-values are in percentage form. R^2s from each time-series regression are reported in percentage form

Regression: $R_{i,t} = \alpha_i + \beta_{i,M}R_{M,t} + \beta_{i,\hat{u}^{DIV}}\hat{u}_t^{DIV} + \beta_{i,\hat{u}^{TERM}}\hat{u}_t^{TERM} + \beta_{i,\hat{u}^{DEF}}\hat{u}_t^{DEF} + \beta_{i,\hat{u}^{RF}}\hat{u}_t^{RF} + \varepsilon_{i,t}$												
	Low	2	3	4	High		Low	2	3	4	High	
	β_{MKT}						$t_{\beta_{MKT}}$					F
Small	1.44	1.23	1.09	1.01	1.02		24.20	22.74	20.76	19.57	18.87	> 100
2	1.44	1.18	1.04	0.98	1.05		31.33	25.11	22.63	21.90	18.76	< 0.01
3	1.38	1.12	0.98	0.90	0.98		39.96	32.34	22.52	21.58	17.66	
4	1.27	1.08	0.97	0.90	0.99		45.46	29.07	24.02	23.95	19.60	
Large	1.01	0.95	0.85	0.78	0.78		42.69	36.55	26.89	20.47	15.34	
	$\beta_{\hat{u}^{DIV}}$						$t_{\beta_{\hat{u}^{DIV}}}$					F
Small	4.75	0.43	−5.02	−5.61	−7.88		0.76	0.08	−0.89	−1.10	−1.44	2.33
2	3.38	−4.01	−7.66	−6.76	−6.51		0.76	−0.79	−1.55	−1.35	−1.09	0.02
3	7.45	−1.30	−5.91	−8.27	−9.18		2.34	−0.35	−1.16	−1.53	−1.36	
4	8.65	−5.83	−6.17	−8.18	−11.81		2.90	−1.29	−1.21	−1.72	−2.04	
Large	−0.78	−3.49	−1.73	−9.69	−9.50		−0.29	−1.18	−0.47	−1.83	−1.49	
	$\beta_{\hat{u}^{TERM}}$						$t_{\beta_{\hat{u}^{TERM}}}$					F
Small	1.51	1.04	1.69	2.82	8.68		0.26	0.26	0.47	0.79	2.24	1.89
2	−8.21	−2.73	−0.19	1.36	5.16		−1.87	−0.75	0.06	0.46	1.44	0.47
3	−6.34	−3.52	−1.72	2.08	4.39		−1.77	−1.17	−0.55	0.55	1.18	
4	−0.73	−1.51	0.21	0.02	2.13		−0.26	−0.59	0.06	0.01	0.54	
Large	−5.98	−3.26	0.78	−0.90	2.90		−2.22	−1.37	0.31	−0.26	0.74	
	$\beta_{\hat{u}^{DEF}}$						$t_{\beta_{\hat{u}^{DEF}}}$					F
Small	−15.45	−14.54	−6.86	−4.79	−8.58		−2.27	−2.17	−1.39	−1.09	−1.68	1.99
2	−10.03	−5.90	−4.78	0.82	−2.20		−2.04	−1.62	−1.37	0.22	−0.49	0.24
3	−11.17	0.22	1.73	4.03	0.81		−2.75	0.08	0.49	1.15	0.18	
4	−5.80	4.81	4.80	8.03	1.08		−2.10	1.92	1.44	2.50	0.25	
Large	−2.45	3.99	9.12	7.25	2.56		−0.96	1.91	3.85	1.91	0.63	
	$\beta_{\hat{u}^{RF}}$						$t_{\beta_{\hat{u}^{RF}}}$					F
Small	4.07	−2.58	0.07	1.03	2.77		0.77	−0.50	0.01	0.22	0.56	1.76
2	−4.37	−5.20	−6.25	−4.57	0.97		−1.00	−1.19	−1.60	−1.16	0.20	1.08
3	−7.63	−4.40	−6.53	−4.08	0.71		−2.29	−1.38	−2.07	−1.09	0.15	
4	−3.43	0.47	−2.04	−5.74	−3.71		−1.12	0.16	−0.69	−1.61	−0.90	
Large	−3.55	−0.59	4.81	−0.89	0.30		−1.14	−0.22	1.41	−0.25	0.06	
				R^2								
				61.51	60.92	63.41	62.41	59.93				
				73.93	73.95	74.47	71.88	67.96				
				79.81	81.80	77.54	73.58	68.96				
				84.99	86.05	80.32	77.51	69.42				
				87.65	86.11	77.89	67.67	55.88				

ables. Kan and Zhang [20] emphasize that checking the joint significance of the assets' factor loadings is an important step in detecting useless factors in the cross-section of returns.

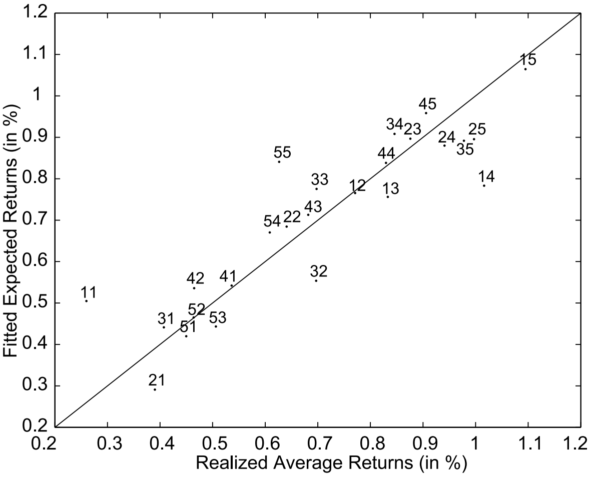
Table 6 contains the results for Eq. (17) which correspond to the second pass of the Fama–MacBeth method. The results reveal that the explanatory power of this model is very close to the one for the Fama–French model re-

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 6

Cross-Sectional Regressions with Loadings on Innovations in State Variables

This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The sample period is from July 1963 to December 2001

The Model with R_M and Innovations in *DIV*, *TERM*, *DEF*, and *RF*							
	γ_0	γ_M	$\gamma_{\hat{u}DIV}$	$\gamma_{\hat{u}TERM}$	$\gamma_{\hat{u}DEF}$	$\gamma_{\hat{u}RF}$	Adj. R^2
Estimate	0.64	−0.07	−1.39	4.89	−0.54	−3.22	77.00
FM *t*-stat	1.74	−0.16	−1.56	4.44	−0.58	−3.79	
SH *t*-stat	1.08	−0.11	−0.99	2.79	−0.37	−2.40	



Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Figure 2

Fitted Expected Returns vs. Average Realized Returns for 1963:07-2001:12.

This figure shows realized average returns (%) on the horizontal axis and fitted expected returns (%) on the vertical axis for 25 size and book-to-market sorted portfolios. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile (1 being the smallest and 5 the largest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For each portfolio, the realized average return is the time-series average of the portfolio return and the fitted expected return is the fitted value for the expected return from the corresponding model. The straight line is the 45-degree line from the origin. The Model with the Excess Market Return and Innovations in the Dividend Yield, Term Spread, Default Spread, and Short-Term T-bill

ported previously in Table 2. Figure 2 plots the fitted versus the realized average returns from the model. It can be seen form the graph that the model based on innovation in predictive variables goes a long way toward explaining the value effect: in general, the fitted expected returns on value portfolios (bigger second digit) are higher than the fitted expected returns on growth portfolios (lower second digit). This is consistent with the data on realized average returns for these portfolios. Further, the model with R_M, $\hat{u}^{DIV}$, $\hat{u}^{TERM}$, $\hat{u}^{DEF}$, and $\hat{u}^{RF}$ is more successful at pricing the portfolios that are challenging for the Fama–French model. The realized returns on growth portfolios within the smallest and largest size groups and the value portfolios within the largest size groups are brought closer to the 45-degree line under the model with the four innovations factors.

In summary, this section has shown that the performance of the model based on innovation in predictive variables is very close to the performance of the Fama–French model in the cross-section of average returns sorted by size and book-to-market. This suggest that the Fama–French factors *HML* and *SMB* might proxy for fundamental state variables that describe variation in investment opportunities over time.

Other Risk-Based Interpretations

Liew and Vassalou [25] show that there is a relation between the Fama–French portfolios *HML* and *SMB* and macroeconomic events. They find that not only in the US but also in several other countries, the corresponding *HML* and *SMB* portfolios contain information about future GDP growth. Therefore, the authors conclude that the size and book-to-market factors are related to future macroeconomic growth. This evidence is consistent with interpreting the *HML* and *SMB* factors as proxies for business cycle risk.

Other studies try to relate the difference in average returns between value and growth portfolios to the time-varying nature of the riskiness of those portfolios. Namely, if value stocks are riskier than growth stocks during bad economic times and if the price of bearing risk is higher during those times, then it follows that value stocks should earn higher average returns than growth stocks. Lettau and Ludvigson [24] document that *HML* is indeed sensitive to bad news in bad macroeconomic times.

Petkova and Zhang [32] is another study that looks are the time-varying risk of value and growth portfolios. They find that the market risk of value stocks is high in bad times when the expected premium for risk is high and it is low in good times when the expected premium for risk is low. What might lead to this time-varying of value and growth stocks? Zhang [35] suggest that the reason might be irreversible investment. He notes that firms with high book-to-market ratios on average will have larger amounts of tangible capital. In addition, it is more costly for firms to reduce than to expand capital. In bad times, firms want to scale down, especially value firms that are less productive than growth firms (Fama and French [13]). Because scaling down is more difficult, value firms are more adversely affected by economic downturns. In good times, growth firms face less flexibility because they tend to invest more. Expanding is less urgent for value firms because their previously unproductive assets have become more productive. In sum, costly reversibility causes value firms to have higher (lower) betas than growth firms in bad (good) times and this contributes to the return differential between these two classes of stocks.

Future Directions

The Fama–French model states that asset returns are driven by three market-wide factors: the excess return on the market portfolio, and the returns on two portfolios related to size (*SMB*) and book-to-market (*HML*). The *HML* and *SMB* portfolios capture the empirical observation that value firms earn higher average returns than growth firms, and small firms earn higher average returns than large firms. The Fama–French model has been very successful at explaining average stock returns, but the exact economic interpretation of the *HML* and *SMB* portfolios has been an issue of debate.

This article examines the risk-based explanation behind the empirical success of the Fama–French model and suggests that the value and size premia arise due to differences in exposure to systematic sources of risk. As mentioned in the introduction, several authors (e. g., Lakonishok, Shleifer, Vishny [22], La Porta, Lakonishok, Shleifer, Vishny [23]), however, claim that the value premium results from irrationality on the side of investors. Namely, investors tend to over-extrapolate recent stock performance: they overvalue the stocks of growth firms and undervalue the stocks of value firms. When the market realizes its mistake, the prices of the former fall, while the prices of the latter rise, resulting in the value premium.

The Fama–French model provides a useful performance benchmark relative to a set of market-wide factors. The results in this article suggest that the Fama–French factors proxy for systematic sources of risk that capture time variation in investment opportunities. However, the debate about the economic interpretation behind the size and value premia is still not settled. Whether they arise as a result of rational compensation for risk or irrational investor behavior is still a matter of controversy.

Bibliography

1. Berk J, Green R, Naik V (1999) Optimal investment, growth options, and security returns. J Finance 54:1553–1608
2. Campbell J (1987) Stock returns and the term structure. J Financial Econ 18:373–399
3. Campbell J (1996) Understanding risk and return. J Political Econ 104:298–345
4. Campbell J, Shiller R (1988) Stock prices, earnings, and expected dividends, J Finance 43:661–676
5. Campbell J, Vuolteenaho T (2004) Bad beta, good beta. Am Econ Rev 5:1249–1275
6. Chan KC, Chen N (1991) Structural and return characteristics of small and large firms. J Finance 46:1467–1484
7. Chen N, Zhang F (1998) Risk and return of value stocks. J Bus 71:501–535
8. Cornell B (1999) Risk, duration, and capital budgeting: New evidence on some old questions. J Bus 72:183–200
9. Daniel K, Titman S (1997) Evidence on the characteristics of cross-sectional variation in stock returns. J Finance 52:1–33
10. Fama E, French K (1989) Business conditions and expected returns on stocks and bonds. J Financial Econ 25:23–49
11. Fama E, French K (1992) The cross-section of expected stock returns. J Finance 47:427–465
12. Fama E, French K (1993) Common risk factors in the returns on bonds and stocks. J Financial Econ 33:3–56
13. Fama E, French K (1995) Size and book-to-market factors in earnings and returns. J Finance 50:131–155
14. Fama E, French K (1996) Multifactor explanations of asset pricing anomalies. J Finance 51:55–84
15. Fama E, MacBeth J (1973) Risk, return, and equilibrium: Empirical tests. J Political Econ 81:607–636
16. Fama E, Schwert W (1977) Asset returns and inflation. J Financial Econ 5:115–146
17. Gomes J, Kogan L, Zhang L (2003) Equilibrium cross-section of returns. J Political Econ 111:693–732
18. Jagannathan R, Wang Z (1996) The conditional CAPM and the cross-section of expected returns. J Finance 51:3–53
19. Jagannathan R, Wang Z (1998) Asymptotic theory for estimating beta pricing models using cross-sectional regressions. J Finance 53:1285–1309
20. Kan R, Zhang C (1999) Two-pass tests of asset pricing models with useless factors. J Finance 54:203–235
21. Kothari SP, Shanken J, Sloan R (1995) Another look at the cross-section of expected returns. J Finance 50:185–224
22. Lakonishok J, Shleifer A, Vishny R (1994) Contrarian investment, extrapolation, and risk. J Finance 49:1541–1578
23. La Porta R, Lakonishok J, Shleifer A, Vishny R (1997) Good news for value stocks: Further evidence on market efficiency. J Finance 52:859–874
24. Lettau M, Ludvigson S (2001) Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. J Political Econ 109:1238–1287

25. Liew J, Vassalou M (2000) Can book-to-market, size, and momentum be risk factors that predict economic growth? J Financial Econ 57:221–245
26. Litterman R, Sheinkman J (1991) Common factors affecting bond returns. J Fixed Income 1:54–61
27. Lo A, MacKinlay C (1990) Data-snooping biases in tests of financial asset pricing models. Rev Financial Stud 3:431–467
28. Long J (1974) Stock prices, inflation, and the term structure of interest rates. J Financial Econ 1:131–170
29. Merton R (1973) An intertemporal capital asset-pricing model. Econometrica 41:867–887
30. Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. Int Econ Rev 25:221–247
31. Petkova R (2006) Do the Fama–French factors proxy for innovations in predictive variables? J Finance 61:581–612
32. Petkova R, Zhang L (2005) Is value riskier than growth? J Financial Econ 78:187–202
33. Shanken J (1992) On the estimation of beta-pricing models. Rev Financial Stud 5:1–34
34. Vassalou M (2003) News related to future GDP growth as a risk factor in equity returns. J Financial Econ 68:47–73
35. Zhang L (2005) The value premium. J Finance 60:67–103

Financial Economics, Time Variation in the Market Return

MARK J. KAMSTRA[1], LISA A. KRAMER[2]
[1] Schulich School of Business, York University, Toronto, Canada
[2] Rotman School of Management, University of Toronto, Toronto, Canada

Article Outline

Glossary

AR(k) An autoregressive process of order k; a time series model allowing for first order dependence; for instance, an AR(1) model is written as $y_t = \alpha + \rho_1 y_{t-1} + \epsilon_t$ where α and ρ are parameters, ρ is typically assumed to be less than 1 in absolute value, and ϵ_t is an innovation term, often assumed to be Gaussian, independent, and identically distributed over t.

ARCH(q) A special case of the GARCH(p, q) model (see below) where $p = 0$.

Basis point A hundredth of one percent.

Bootstrap A computer intensive resampling procedure, where random draws with replacement from an original sample are used, for instance to perform inference.

Discount rate The rate of return used to discount future cashflows, typically calculated as a risk-free rate (e. g. the 90-day US T-bill rate) plus an equity risk premium.

Equity premium puzzle The empirical observation that the ex post equity premium (see entry below) is higher than is indicated by financial theory.

Ex ante equity premium The extra return investors *expect* they will receive for holding risky assets, over and above the return they would receive for holding a risk-free asset like a Treasury bill. "Ex ante" refers to the fact that the expectation is formed in advance.

Ex post equity premium The extra return investors received *after* having held a risky asset for some period of time. The ex post equity premium often differs from the ex ante equity premium due to random events that impact a risky asset's return.

Free cash flows Cash flows that could be withdrawn from a firm without lowering the firm's current rate of growth. Free cash flows are substantially different from accounting earnings and even accounting measures of the cash flow of a firm.

Fundamental valuation The practice of determining a stock's intrinsic value by discounting cash flows to their present value using the required rate of return.

GARCH(p, q) Generalized autoregressive conditional heteroskedasticity of order (p, q), where p is the order of the lagged variance terms and q is the order of the lagged squared error terms; a time series model allowing for dependence in the conditional variance of a random variable, y. A GARCH(1,1) model is specified as:

$$y_t = \alpha + \epsilon_t ; \quad \epsilon_t \sim \left(0, h_t^2\right)$$
$$h_t^2 = \theta + \beta h_{t-1}^2 + \gamma \epsilon_{t-1}^2 ,$$

where α, θ, β, and γ are parameters and ϵ_t is an innovation term.

Market anomalies Empirical regularities in financial market prices or returns that are difficult to reconcile with conventional theories and/or valuation methods.

Markov model A model of a probabilistic process where the random variable can only take on a finite number of different values, typically called states.

Method of moments A technique for estimating parameters (like parameters of the conditional mean and conditional variance) by matching sample moments, then solving the equations for the parameters to be estimated.

SAD Seasonal Affective Disorder, a medical condition by which reduced daylight in the fall and winter leads to seasonal depression for roughly ten percent of the world's population.

Sensation seeking A measure used by psychologists to capture an individual's degree of risk tolerance. High sensation-seeking tendency correlates with low risk tolerance, including tolerance for risk of a financial nature.

Simulated method of moments A modified version of the method of moments (see entry above) that is based on Monte Carlo simulation, used in situations when the computation of analytic solutions is infeasible.

Definition of the Subject

The realized return to any given asset varies over time, occasionally in a dramatic fashion. The value of an asset, its *expected* return, and its volatility, are of great interest to investors and to policy makers. An asset's expected return in excess of the return to a riskless asset (such as a short-term US Treasury bill) is termed the equity premium. The value of the equity premium is central to the valuation of risky assets, and hence a much effort has been devoted to determining the value of the equity premium, whether it varies, and if it varies, how predictable it is. Any evidence of predictable returns is either evidence of a predictably varying equity premium (say, because risk varies predictably) or a challenge to the rationality of markets and the efficient allocation of our society's scarce resources.

In this article, we start by considering the topic of valuation, with emphasis on simulation-based techniques. We consider the valuation of income-generating assets in the context of a constant equity premium, and we also explore the consequences of allowing some time-variation and predictability in the equity premium. Next we consider the equity premium puzzle, discussing a simulation-based technique which allows for precise estimation of the value of the equity premium, and which suggests some constraints on the types of models that should be used for specifying the equity premium process. Finally, we focus on evidence of seasonally varying expected returns in financial markets. We consider evidence that as a whole either presents some challenges to traditional hypotheses of efficient markets, or suggests agents' risk tolerance may vary over time.

Introduction

The pricing of a firm is conceptually straightforward. One approach to valuing a firm is to use historical dividend payments and discount rate data to forecast future payments and discount rates. Restrictions on the dividend and discount rate processes are typically imposed to produce an analytic solution to the fundamental valuation equation (an equation that involves calculating the expectation of an infinite sum of discounted dividends).

Common among many of the available valuation techniques is some form of consideration of multiple scenarios, including good and bad growth and discount rate evolutions, with valuation based on a weighted average of prices from the various scenarios. The valuation technique we focus some attention on, the Donaldson and Kamstra [14] (henceforth DK) methodology, is similar to pricing path-dependent options, as it utilizes Monte Carlo simulation techniques and numerical integration of the possible paths followed by the joint processes of dividend growth and discount rates, explicitly allowing path-dependence of the evolutions. The DK method is very similar in spirit to other approaches in the valuation literature which consider multiple scenarios. One distinguishing feature of the DK methodology we consider is the technique it employs for modeling the discount rate.

Cochrane [9] highlights three interesting approaches for modeling the discount rate: a constant discount rate, a consumption-based discount rate, and a discount rate equal to some variable reference return plus a risk premium. Virtually the entire valuation literature limits its attention to the constant discount rate case, as constant discount rates lead to closed-form solutions to many valuation formulas. DK explore all three methods for modeling the discount rate and find they lead to qualitatively similar results. However, their quantitative results indicate an overall better fit to the price and return data when using a reference return plus a risk premium. Given DK's findings, we use a discount rate equal to some variable reference return plus a risk premium. In implementing this approach for modeling the discount rate used in valuation, it is simplest to assume a *constant* equity premium is added to the reference rate, in particular since the reference rate is permitted to vary (since it is typically proxied using a variable rate like the three-month US T-bill rate). We do not, however, restrict ourselves to the constant equity premium case.

Using the simulation-based valuation methodology of DK and the method of simulated moments, we explore the evidence for a time-varying equity premium and its implications for a long-standing puzzle in financial economics, the equity premium puzzle of Mehra and Prescott [51]. Over the past century the average annual return to investing in the US stock market has been roughly 6% higher than the return to investing in risk-free US T-bills. Making use of consumption-based asset-pricing models, Mehra and Prescott argue that consumption within the US has not been sufficiently volatile to warrant such a large premium on risky stocks relative to riskless bonds, leading them to describe this large premium as the "equity premium puzzle."

Utilizing simulations of the distribution from which ex post equity premia are drawn, conditional on various possible values for investors' ex ante equity premium and calibrated to S&P 500 dividends and US interest rates, we present statistical tests that show a true ex ante equity premium as low as 2% could easily produce ex post premia of 6%. This result is consistent with the well-known observation that ex post equity premia are observed with error, and a large range of realized equity premia are consistent with any given value of the ex ante equity premium. Examining the marginal and joint distributions of financial statistics like price-dividend ratios and return volatility that arise in the simulations versus actual realizations from the US economy, we argue that the range of ex ante equity premia most consistent with the US market data is very close to 3.5%, and the ex ante equity premium process is very unlikely to be constant over time.

A natural question to ask is why might the equity premium fluctuate over time? There are only two likely explanations: changing risk or changing risk aversion. Evidence from the asset-pricing literature, including [20,37,49], and many others shows that priced risk varies over time. We explore some evidence that risk aversion itself may vary over time, as revealed in what is often termed market anomalies. Market anomalies are variations in expected returns which appear to be incongruous with variations in discount rates or risk. The most stark anomalies have to do with deterministic asset return seasonalities, including seasonalities at the weekly frequency such as the weekend effect (below-average equity returns on Mondays), annual effects like the above-average equity returns typically witnessed in the month of January, and other effects like the lower-than-average equity returns often witnessed following daylight saving time-change weekends, and opposing cyclicality in bond versus equity returns correlated to the length of day (known as the SAD effect). We briefly review some of these outstanding puzzles, focusing our attention on the SAD effect and the daylight saving effect.

Valuation

Overview

We begin our discussion of valuation with a broad survey of the literature, including dividend-based valuation, relative valuation, and accounting-based methods. We introduce dividend-based valuation first.

Fundamental valuation techniques that utilize dividends in a *discrete* time framework include Gordon [25], Hawkins [30], Michaud and Davis [53], Farrell [22], Sorensen and Williamson [73], Rappaport [63], Barsky and DeLong [2], Hurley and Johnson [33], [34], Donaldson and Kamstra [14], and Yao [78]. Invariably these approaches are partial equilibrium solutions to the valuation exercise. Papers that use *continuous* time tools to evaluate the fundamental present value equation include Campbell and Kyle [6], Chiang, Davidson, and Okuney [8], Dong and Hirshleifer [17], and Bakshi and Chen [3]. The Dong and Hirshleifer [17] and Bakshi and Chen [3] papers conduct valuation by assuming dividends are proportional to earnings and then modeling earnings. Continuous time papers in this literature typically start with the representative agent/complete markets economic paradigm. Models are derived from primitive assumptions on markets and preferences, such as the equilibrium condition that there exist no arbitrage opportunities, dividend (cash flow) growth rates follow an Ornstein–Uhlenbeck mean-reverting process, and preferences over consumption are represented by the log utility function. Time-varying stochastic discount rates (i. e. the pricing kernel) fall out of the marginal rate of utility of consumption in these models, and the solution to the fundamental valuation problem is derived with the same tools used to price financial derivatives. A critique of dividend-discounting methods is that dividends are typically smoothed and are set low enough so that the dividend payments can be maintained through economic downturns. Authors such as Hackel and Livnat (see p. 9 in [27]) argue that these sorts of considerations imply that historical records of dividend payments may therefore be poor indicators of future cash payments to investors.

A distinct valuation approach, popular amongst practitioners, determines the value of inactively traded firms by finding an actively traded firm that has similar risk, profitability, and investment-opportunity characteristics and then multiplying the actively traded firm's price-earnings (P/E) ratio by the inactively traded firm's earnings. This approach to valuation is often referred to as the relative value method or the constant P/E model. References to this sort of approach can be found in textbooks like [4], and journal articles such as [60,62].

There are also several valuation approaches that are based on the book value of equity, abnormal earnings, and free-cash flows. These approaches are linked to dividends and hence to formal fundamental valuation by well-established accounting relationships. They produce price estimates by valuing firm assets and income streams. The most popular of this class of techniques include the residual income and free-cash-flow methods. See [23,57,61] for further information. All of these valuation methods implicitly or explicitly take the present value of the stream of firm-issued dividends to the investor. The motivation for considering accounting relationships is that these accounting measures are not easily manipulated by firms and so should reflect more accurately the ability of firms to generate cashflows and hence allow more accurate assessments of the fundamental value of a firm than techniques based on dividends.

Fundamental Valuation Methods in Detail

Now that we have surveyed the valuation literature in general, we turn to a formal derivation of several *fundamental* valuation techniques. Investor rationality requires that the current market price P_t of a stock which will pay a per share dividend (cash payment) D_{t+1} one period from now and then sell for P_{t+1}, discounting payments received during period t (i. e., from the beginning of period t to the beginning of period $t+1$) at rate r_t, must satisfy Eq. (1):

$$P_t = \mathcal{E}_t \left\{ \frac{P_{t+1} + D_{t+1}}{1 + r_t} \right\} . \qquad (1)$$

$\mathcal{E}_t$ is the expectations operator conditional on information available up to the end of period t. Solving Eq. (1) forward under the transversality condition that the expected present value of P_{t+k} goes to zero as k goes to infinity (a "no-bubble" assumption) produces the familiar result that the market price equals the expected present value of future dividends (cash payments); i. e.,

$$P_t = \sum_{k=0}^{\infty} \mathcal{E}_t \left\{ \left(\prod_{i=0}^{k} \left[\frac{1}{1 + r_{t+i}} \right] \right) D_{t+k+1} \right\} . \qquad (2)$$

Defining the growth rate of dividends from the beginning of period t to the beginning of period $t+1$ as $g_t^{\mathrm{d}} \equiv (D_{t+1} - D_t)/D_t$ it follows that

$$P_t = D_t \mathcal{E}_t \left\{ \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k} \left[\frac{1 + g_{t+i}^{\mathrm{d}}}{1 + r_{t+i}} \right] \right) \right\} . \qquad (3)$$

Equation (3) is the fundamental valuation equation, which is not controversial and can be derived under the law of one price and non-satiation alone, as by Rubinstein [69] and others. Notice that the cash payments D_{t+k} include all cash disbursements from the firm, including cash dividends and share repurchases. Fundamental valuation methods based directly on Eq. (3) are typically called dividend discount models.

Perhaps the most famous valuation estimate based on Eq. (3) comes from the Gordon [25] Growth Model. If dividend growth rates and discount rates are constant, then it is straightforward to derive the Gordon fundamental price estimate from Eq. (3):

$$P_t^{\mathrm{G}} = D_t \left[\frac{1 + g^{\mathrm{d}}}{r - g^{\mathrm{d}}} \right] , \qquad (4)$$

where r is the constant discount rate value and g^{d} is the (conditionally) constant growth rate of dividends. To produce the Gordon Growth Model valuation estimate, all we need are estimates of the dividend growth rate and discount rate, which can be obtained in a variety of ways, including the use of historically observed dividends and returns.

Extensions of the Gordon Growth Model exploit the fundamental valuation equation, imposing less stringent assumptions. The simple Gordon Growth Model imposes a constant growth rate on dividends (dividends are expected to grow at the same rate every period) while Hurley and Johnson [33] and [34] and Yao [78] develop Markov models (models that presume a fixed probability of, say, maintaining the dividend payment at current levels, and a probability of raising it, thus incorporating more realistic dividend growth processes). Two examples of these models found in Yao [78] are the Additive Markov Gordon model (Eq. (1) of Yao [78] and the Geometric Markov Gordon model (Eq. (2) of Yao [78]). These models can be interpreted as considering different scenarios for dividend growth for a particular asset, estimating the appropriate price for the asset under each scenario, and then averaging the prices using as weights the probability of given scenarios being observed.

The Additive Markov Gordon Growth Model is:

$$P_t^{\mathrm{ADD}} = D_t/r + \left[1/r + (1/r)^2 \right] \left(q^{\mathrm{u}} - q^{\mathrm{d}} \right) \Delta , \qquad (5)$$

where r is the average discount rate, q^{u} is the proportion of the time the dividend increases, q^{d} is the proportion of the time the dividend decreases, and $\Delta = \sum_{t=2}^{T} |D_t - D_{t-1}|/(T-1)$ is the average absolute value of the level change in the dividend payment.

The Geometric Markov Gordon Growth Model is:

$$P_t^{\mathrm{GEO}} = D_t \left[\frac{1 + (q^{\mathrm{u}} - q^{\mathrm{d}})\Delta^{\%}}{r - (q^{\mathrm{u}} - q^{\mathrm{d}})\Delta^{\%}} \right] , \qquad (6)$$

where $\Delta^{\%} = \sum_{t=2}^{T} |(D_t - D_{t-1})/D_{t-1}|/(T-1)$ is the average absolute value of the percentage rate of change in the dividend payment.

The method of DK is also an extension of the Gordon Growth Model, taking the discounted dividend growth model of Eq. (3) and re-writing it as

$$P_t = D_t \sum_{k=0}^{\infty} \mathcal{E}_t \left\{ \prod_{i=0}^{k} y_{t+i} \right\} , \qquad (7)$$

where $y_{t+i} = (1 + g_{t+i}^{\mathrm{d}})/(1 + r_{t+i})$ is the discounted dividend growth rate. Under the DK method, the fundamental price is calculated by forecasting the range of possible evolutions of y_{t+i} up to some distant point in the future, period $t + I$, calculating $PV = D_t \sum_{k=0}^{I} (\prod_{i=0}^{k} y_{t+i})$ for each possible evolution of y_{t+i}, and averaging these values of PV across all the possible evolutions. (The value of I is chosen to produce a very small truncation error. Values of $I = 400$ to 500 for annual data have been found by DK to suffice). In this way, the DK approach mirrors other extensions of the Gordon Growth Model. It is primarily distinguished from other approaches that extend the Gordon

Growth Model in two regards. First, more sophisticated time series models, estimated with historical data, are used to generate the different outcomes (scenarios) by application of Monte Carlo simulation. Second, in contrast to typical modeling in which only dividend growth rates vary, the joint evolution of cashflow growth rates and discount rates are explicitly modeled as time-varying.

Among the attractive features of the free-cash-flow and residual income valuation methods is that they avoid the problem of forecasting dividends, by exploiting relationships between accounting data and dividends. It is the practical problem of forecasting dividends to infinity that have led many researchers to explore methods based on accounting data. See, for instance, Penman and Sougiannis [61].

Assume a flat term structure (i. e., a constant discount rate $r_t = r$ for all t) and write

$$P_t = \sum_{k=1}^{\infty} \frac{\mathcal{E}_t\{D_{t+k}\}}{(1+r)^k} . \tag{8}$$

The clean-surplus relationship relating dividends to earnings is invoked in order to derive the residual income model:

$$B_{t+k} = B_{t+k-1} + E_{t+k} - D_{t+k} , \tag{9}$$

where B_{t+k} is book value and E_{t+k} is earnings per share. Solving for D_{t+k} in Eq. (9) and substituting into Eq. (8) yields

$$P_t = \sum_{k=1}^{\infty} \frac{\mathcal{E}_t\{B_{t+k-1} + E_{t+k} - B_{t+k}\}}{(1+r)^k} ,$$

or

$$\begin{aligned} P_t &= B_t + \sum_{k=1}^{\infty} \frac{\mathcal{E}_t\{E_{t+k} - r \cdot B_{t+k-1}\}}{(1+r)^k} - \frac{\mathcal{E}_t\{B_{t+\infty}\}}{(1+r)^{\infty}} \\ &= B_t + \sum_{k=1}^{\infty} \frac{\mathcal{E}_t\{E_{t+k} - r \cdot B_{t+k-1}\}}{(1+r)^k} , \end{aligned} \tag{10}$$

where $B_{t+\infty}/(1+r)^{\infty}$ is assumed to equal zero. $E_{t+k} - r \cdot B_{t+k-1}$ is termed abnormal earnings.

To derive the free cash flow valuation model, we relate dividends to cash flows with a financial assets relation in place of the clean surplus relation:

$$fa_{t+k} = fa_{t+k-1} + i_{t+k} + c_{t+k} - D_{t+k} , \tag{11}$$

where fa_{t+k} is financial assets net of financial obligations, i_{t+k} is interest revenues net of interest expenses, and c_{t+k} is cash flows realized from operating activities net of investments in operating activities, all of which can be positive or negative. A net interest relation is often assumed,

$$i_{t+k} = r fa_{t+k-1} . \tag{12}$$

See Fetham and Ohlson [23] for further discussion. Solving for D_{t+k} in Eq. (11) and substituting into Eq. (8), utilizing Eq. (12) and assuming the discounted present value of financial assets fa_{t+k} goes to zero as k increases, yields the free-cash-flow valuation equation:

$$P_t = fa_t + \sum_{k=1}^{\infty} \frac{\mathcal{E}_t\{c_{t+k}\}}{(1+r)^k} . \tag{13}$$

More on the Fundamental Valuation Method of Donaldson and Kamstra

A number of approaches can be taken to conduct valuation using the DK model shown in Eq. (7). By imposing a very simple structure for the conditional expectation of discounted dividend growth rate (y_t in Eq. (7)), the expression can be solved analytically, for instance by assuming that the discounted dividend growth rate is a constant. As shown by DK, however, analytic solutions become complex for even simple ARMA models, and with sufficient non-linearity, the analytics can be intractable. For this reason, we present a general solution algorithm based on the DK method of Monte Carlo simulation.

This method simulates y_t into the future and performs a numerical (Monte Carlo) integration to estimate the terms $\{\prod_{k=0}^{i} y_{t+k}\}$ where $y_{t+k} = (1 + g^{\mathrm{d}}_{t+k})/(1 + r_{t+k})$ in the classic case of a dividend-paying firm. A general heuristic is as follows:

Step I: Model y_t, $t = 1, \ldots, T$, as conditionally time-varying, for instance as an AR(k)-GARCH(p, q) process, and use the estimated model to make conditional mean forecasts $\hat{y}_t$, $t = 1, \ldots, T$, and variance forecasts, conditional on data observed only before period t. Ensure that this model is consistent with theory, for instance that the mean level of y is less than one. This mean value can be calibrated to available data, such as the mean annual y value of 0.94 observed in the last 50 years of S&P 500 data. Recall that although analytic solutions are available for simple processes, the algorithm presented here is applicable to virtually arbitrarily non-linear conditional processes for the discounted cash payment rate y.

Step IIa: Simulate discounted cash payment growth rates. That is, produce ys that might be observed in period t given what is known at period $t-1$. To do this for a given period t, simulate a population of J independent possible shocks (say draws from a normal distribution with mean zero and appropriate variance, or bootstrapped from the data) $\epsilon_{t,j}$, $j = 1, \dots, J$, and add these shocks separately to the conditional mean forecast $\hat{y}_t$ from Step I, producing $y_{t,j} = \hat{y}_t + \epsilon_{t,j}$, $j = 1, \dots, J$. The result is a simulated cross-section of J possible realizations of y_t standing at time $t-1$, i. e. different paths the economy may take next period.

Step IIb: Use the estimated model from Step I to make the conditional mean forecast $\hat{y}_{t+1,j}$, conditional on only the jth realization for period t (i. e., $y_{t,j}$ and $\epsilon_{t,j}$) and the data known at period $t-1$, to form $y_{t+1,j}$.

Step IIc: Repeat Step IIb to form $y_{t+2,j}, y_{t+3,j}, \dots, y_{t+I,j}$ for each of the J economies, where I is the number of periods into the future at which the simulation is truncated. Form the perfect foresight present value ($P^*_{t,j}$) for each of the J possible economies:

$$P^*_{t,j} = A_t\Big(y_{t,j} + y_{t,j}y_{t+1,j} + y_{t,j}y_{t+1,j}y_{t+2,j} + \dots + \prod_{i=0}^{I} y_{t+i,j}\Big)\,; \quad j = 1, \dots, J\,.$$

Provided I is chosen to be large enough, the truncated terms $\prod_{i=0}^{K} y_{t+i,j}$, $K = I+1, \dots, \infty$ will be negligible.

Step III: Calculate the DK fundamental price for each $t = 1, \dots, T$:

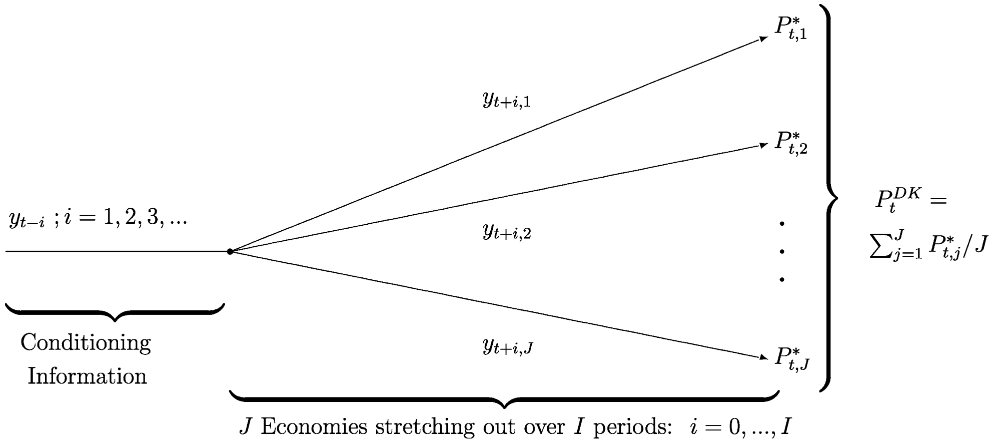
$$P^{\mathrm{DK}}_t = \sum_{j=1}^{J} P^*_{t,j}/J\,. \qquad (14)$$

These fundamental price estimates P^{DK}_t can be compared to the actual price (if market prices exist) at the beginning of period t to test for bubbles as demonstrated by DK, or if period t is the future, P^{DK}_t is the fundamental price forecast. This procedure is represented diagrammatically in Exhibit 1.

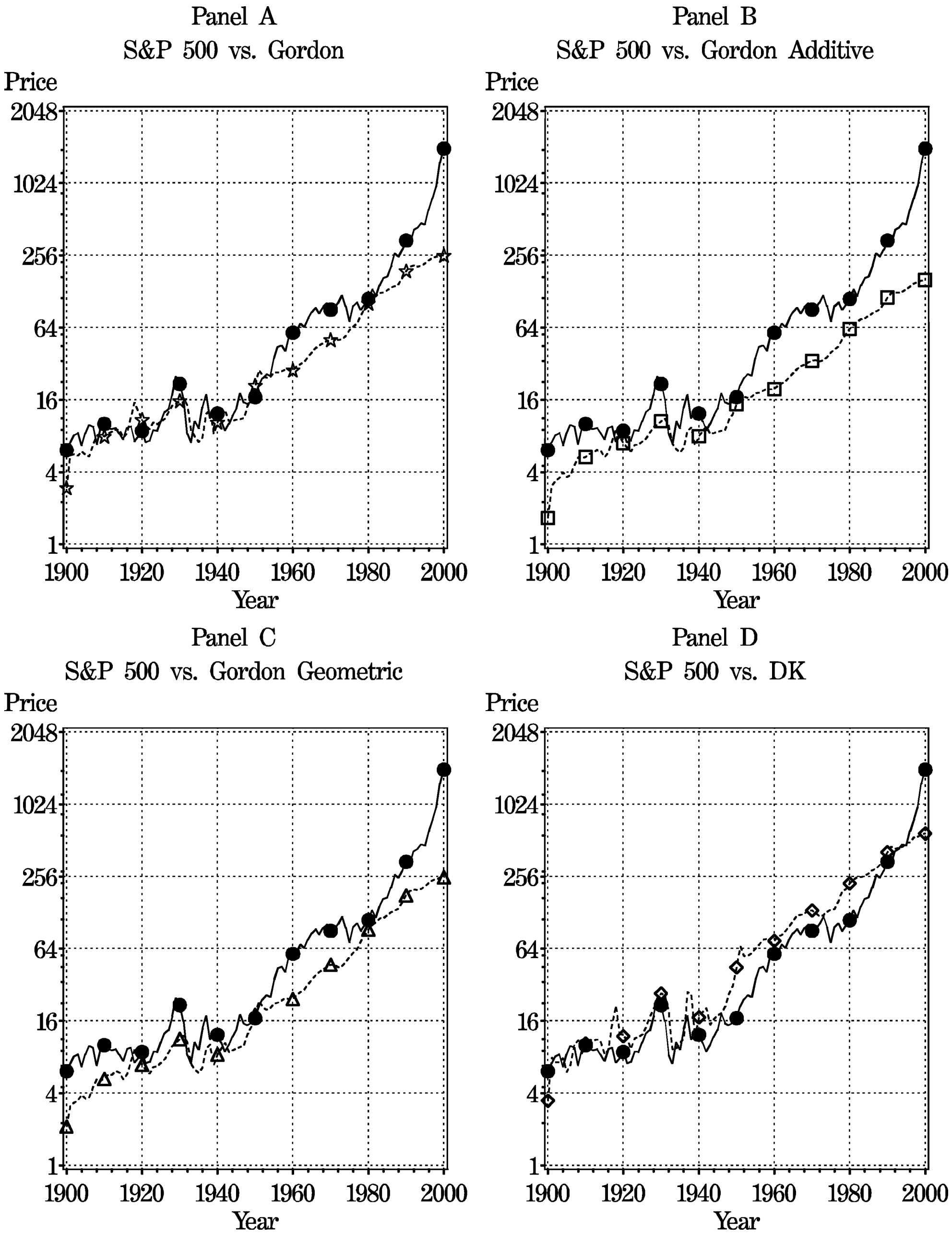
To illustrate the sort of forecasts that can be produced using this technique, we illustrate graphically the S&P 500 index over the past 100 years together with predicted values based on the Gordon Growth Model and the DK method. The free-cash-flow and residual income methods are not easily adapted to forecasting index prices like the S&P 500, and so are omitted here. The type of data depicted in the following figure is described in some detail by Kamstra [39].

Figure 1 has four panels. In the panels, we plot the level of the S&P 500 index (marked with bullets and a solid line) alongside price forecasts from each of the valuation techniques. In Panel A we plot the index together with the basic Gordon Growth Model price forecasts (marked with stars), in Panels B and C we plot the index together with the Additive and Geometric Gordon Growth Models' forecasts (with squares and triangles respectively), and in Panel D we plot the index alongside the DK method's forecasts (marked with diamonds). In each panel the price scale is logarithmic.

We see in Panels A, B, and C that the use of the any of the Gordon models for forming annual forecasts of the S&P 500 index level produces excessively smooth price forecasts. (If we had plotted return volatility, then the market returns would appear excessively volatile in comparison to to forecasted returns). Evidence of periods of inflated market prices relative to the forecasted prices, i. e.,


$P^*_{t,1}$
$y_{t+i,1}$
$P^*_{t,2}$
y_{t-i} ; $i = 1, 2, 3, \dots$
$y_{t+i,2}$
$P^{DK}_t = \sum_{j=1}^{J} P^*_{t,j}/J$
Conditioning Information
$y_{t+i,J}$
$P^*_{t,J}$
J Economies stretching out over I periods: $i = 0, \dots, I$

Financial Economics, Time Variation in the Market Return, Exhibit 1
Diagram of DK Monte Carlo integration



Financial Economics, Time Variation in the Market Return, Figure 1
S&P 500 index level versus price forecasts from four models. S&P 500 index: ●, Gordon Growth price: ★, Additive Gordon Growth price: □, Geometric Gordon Growth price: △, DK price ◇

evidence of price bubbles, is apparent in the periods covering the 1920s, the 1960s, the last half of the 1980s, and the 1990s. However, if the Gordon models are too simple (since each Gordon-based model ignores the forecastable nature of discount rates and dividend growth rates), then this evidence may be misleading.

In Panel D, we see that the DK model is better able to capture the volatility of the market, including the boom of the 1920s, the 1960s and the 1980s. The relatively better performance of the DK price estimate highlights the importance of accounting for the slow fade rate of dividend growth rates and discount rates, i. e., the autocorrelation of these series. The failure of the DK method to capture the height of the 1990s boom leaves evidence of surprisingly high prices during the late 1990s. If the equity premium fell in the 1990s, as some researchers have speculated (see for instance Pástor and Stambaugh [59]), then all four sets of the plotted fundamental valuation forecasts would be expected to produce forecasts that undershoot actual prices in the 1990s, as all these methods incorporate a constant equity premium. If this premium were set too high, future cashflows would be discounted too aggressively, biasing the valuation methods downward.

The Equity Premium Puzzle

The fact that all four fundamental valuation methods we consider spectacularly fail to capture the price boom of the 1990s, possibly as a result of not allowing a time-varying equity premium, sets the stage to investigate the equity premium puzzle of Mehra and Prescott [51]. The equity premium is the extra return, or premium, that investors demand in order to be compelled to purchase risky stock instead of risk-free debt. We call this premium the ex ante equity premium (denoted π_{e}), and it is formally defined as the difference between the expected return on risky assets, $\mathcal{E}\{R\}$, and the expected risk-free rate, $\mathcal{E}\{r_{\text{f}}\}$:

$$\pi_{\text{e}} \equiv \mathcal{E}\{R\} - \mathcal{E}\{r_{\text{f}}\} \ . \tag{15}$$

The ex post equity premium is typically estimated using historical equity returns and risk-free rates, as we do not observe the ex ante premium. Define $\overline{R}$ as the average historical annual return on the S&P 500 and $\overline{r}_{\text{f}}$ as the average historical return on US T-bills. A standard approach to calculate ex post equity premium, $\hat{\pi}_{\text{e}}$, is:

$$\hat{\pi}_{\text{e}} \equiv \overline{R} - \overline{r}_{\text{f}} \, . \tag{16}$$

Of course it is unlikely that the stock return we estimate ex post equals investors' anticipated ex ante return. Thus a 6% ex post equity premium in the US data may not be a challenge to economic theory. The question we ask is therefore: if investors' true ex ante premium is X%, what is the probability that the US economy could randomly produce an ex post premium of at least 6%? We can then argue whether or not the 6% ex post premium observed in the US data is consistent with various ex ante premium values, X%, with which standard economic theory may be more compatible. We can also consider key financial statistics and yields from the US economy to investigate if an X% ex ante equity premium could likely be consistent with the combinations that have been observed, such as high Sharpe ratio and low dividend yields, low interest rates and high ex post equity premia, and so on.

Authors have investigated the extent to which ex ante considerations may impact the realized equity premium. For example, Rietz [65] investigated the effect that the fear of a serious, but never realized, depression would have on equilibrium asset prices and equity premia. Jorion and Goetzmann [38] take the approach of comparing the US stock market's performance with stock market experiences in many other countries. They find that, while some markets such as the US and Canada have done very well over the past century, other countries have not been so fortunate; average stock market returns from 1921 to 1996 in France, Belgium, and Italy, for example, are all close to zero, while countries such as Spain, Greece, and Romania have experienced negative returns. It is difficult, however, to conduct statistical tests because, first, the stock indices Jorion and Goetzmann consider are largely contemporaneous and returns from the various indices are not independent. Statistical tests would have to take into account the panel nature of the data and explicitly model covariances across countries. Second, many countries in the comparison pool are difficult to compare directly to the United States in terms of economic history and underlying data generating processes. (Economies like Egypt and Romania, for example may have equity premia generated from data generating processes that differ substantially from that of the US).

There are some recent papers that make use of fundamental information in examining the equity premium. One such paper, Fama and French [21], uses historical dividend yields and other fundamental information to calculate estimates of the equity premium which are smaller than previous estimates. Fama and French obtain point estimates of the ex post equity premium ranging from 2.55% (based on dividend growth rate fundamentals) to 4.78% (based on bias-adjusted earnings growth rate fundamentals), however these estimates have large standard errors. For example, for their point estimate of 4.32% based on non-bias-adjusted earnings growth rates, a 99% confidence interval stretches from approximately −1% to about 9%. Mehra and Prescott's [51] initially troubling estimate of 6% is easily within this confidence interval and is in fact within one standard deviation of the Fama and French point estimate.

Calibrating to economy-wide dividends and discount rates, Donaldson, Kamstra, and Kramer [16] employ simulation methods similar to DK to simulate a distribution of possible price and return outcomes. Comparing these simulated distributions with moments of the actual data then permits them to test various models for the equity premium process. Could a realized equity premium of 6% be consistent with an ex ante equity premium of 2%?

Could an ex ante equity premium of 2% have produced the low dividend yields, high ex post equity premia, and high Sharpe ratios observed in the US over the last half century?

A summary of the basic methodology implemented by Donaldson, Kamstra, and Kramer [16], is as follows:

(a) Assume a mean value for the equity premium that investors demand when they first purchase stock (e. g., 2%) and a time series process for the premium, say a deterministic drift downward in the premium of 5 basis points per year, asymptoting no lower than perhaps 1%. This assumed premium is added to the risk-free interest rate to determine the discount rate that an investor would rationally apply to a forecasted dividend stream in order to calculate the present value of dividend-paying stock.

(b) Estimate econometric models for the time-series processes driving dividends and interest rates in the US economy (and, if necessary, for the equity premium process), allowing for autocorrelation and covariation. Then use these models to Monte Carlo simulate a variety of potential paths for US dividends, interest rates, and equity premia. The simulated paths are of course different in each of these simulated economies because different sequences of random innovations are applied to the common stochastic processes in each case. However, the key drivers of the simulated economies themselves are all still identical to those of the US economy since all economies share common stochastic processes fitted to US data.

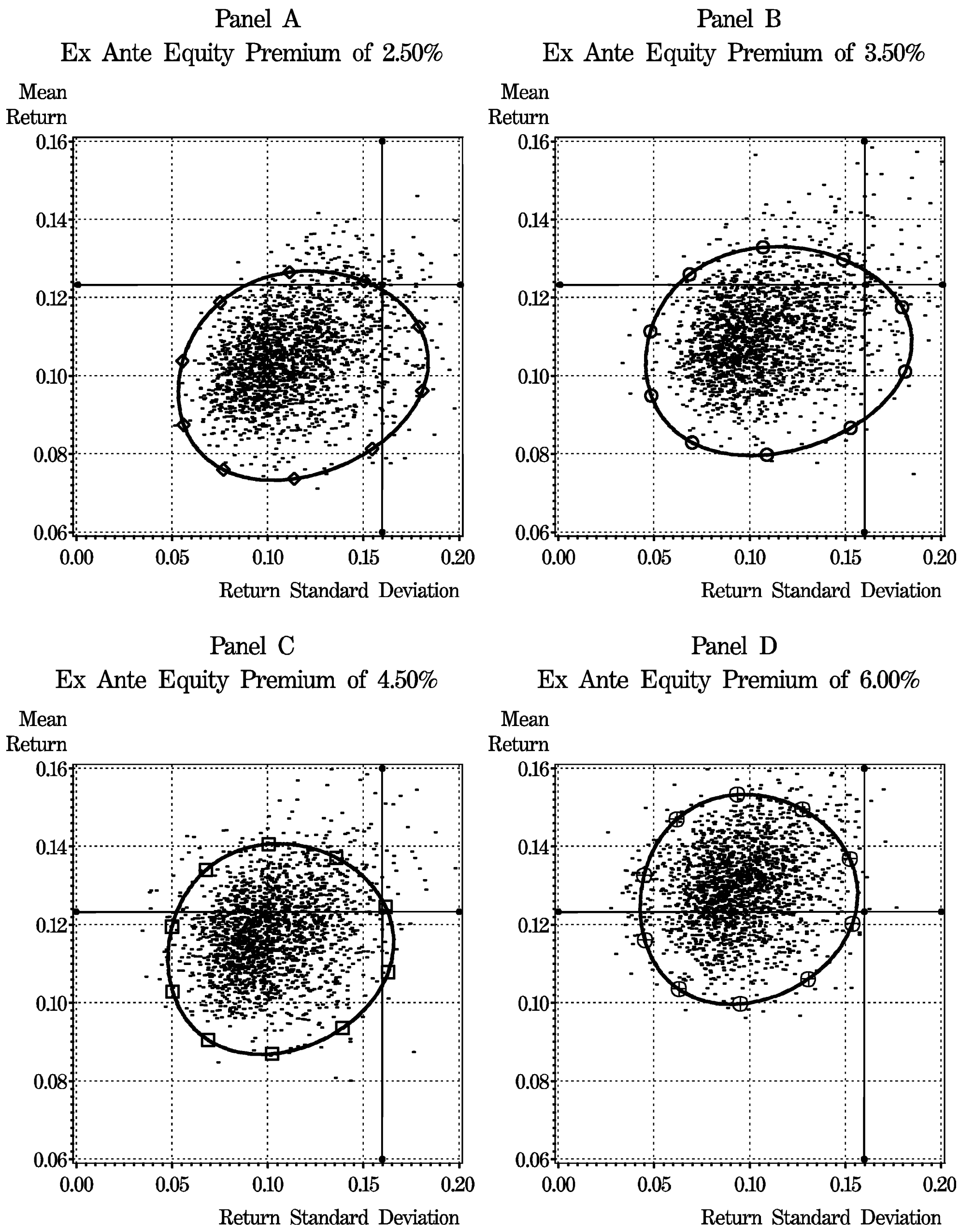
(c) Given the assumed process for the equity premium investors demand ex ante (which is the same for all simulated economies in a given experiment), use a discounted-dividend model to calculate the fundamental stock returns (and hence ex post equity premia) that arise in each simulated economy. All economies have the same ex ante equity premium process, and yet all economies have different ex post equity premia. Given the returns and ex post equity premia for each economy, as well as the means of the interest rates and dividend growth rates produced for each economy, it is feasible to calculate various other important characteristics, like Sharpe ratios and dividend yields.

(d) Examine the distribution of ex post equity premia, interest rates, dividend growth rates, Sharpe ratios, and dividend yields that arise conditional on various values of the ex ante equity premia. Comparing the performance of the US economy with intersections of the various univariate and multivariate distributions of these quantities and conducting joint hypothesis tests allows the determination of a narrow range of equity premia consistent with the US market data. Note that this is the method of simulated moments, which is well adapted to estimate the ex ante equity premium. The simulated method of moments was developed by McFadden [50] and Pakes and Pollard [58]. Duffie and Singleton [18] and Corradi and Swanson [11] employ simulated method of moments in an asset pricing context.

Further details on the simulation methodology are provided by Donaldson, Kamstra, and Kramer [16]. They make use of annual US stock and Treasury data observed from 1952 through 2004, with the starting year of 1952 motivated by the US Federal Reserve Board's adoption of a modern monetary policy regime in 1951. The model that generated the data we use to illustrate this simulation methodology is Model 6 of Donaldson, Kamstra, and Kramer [16], a model that allows for trending, autocorrelated, and co-varying dividend growth rates, interest rates and equity premia, as well as for a structural break in the equity premium process. We show later that allowing for trends and structural breaks in the equity premium process is a crucial factor in the model's ability to capture the behavior of the observed US market data.

We focus on the intuition behind the Donaldson, Kamstra, and Kramer technique by looking at bivariate plots of simulated data, conditional on various values of the ex ante equity premium. In every case, the pair of statistics we plot are dependent on each other in some way, allowing us to make interesting conditional statements. Among the bivariate distributions we consider, we will see some that serve primarily to confirm the ability of our simulations to produce the character and diversity of results observed in US markets. Some sets of figures rule out ex ante equity premia below 2.5% while others rule out ex ante equity premia above 4.5%. Viewed collectively, the figures serve to confirm that the range of ex ante equity premia consistent with US market data is in the close vicinity of 3.5%.

Figure 2 contains joint distributions of mean returns and return standard deviations arising in our simulations based on four particular values of the ex ante equity premium (2.5% in Panel A, 3.5% in Panel B, 4.5% in Panel C, and 6% in Panel D). Each panel contains a scatter plot of two thousand points, with each point representing a pair of statistics (mean return and return standard deviation) arising in one of the simulated half-century economies. The combination based on the US realization is shown in each plot with a crosshair (a pair of solid straight lines with the intersection marked by a solid dot). The set of


Panel A
Ex Ante Equity Premium of 2.50%

Panel B
Ex Ante Equity Premium of 3.50%

Panel C
Ex Ante Equity Premium of 4.50%

Panel D
Ex Ante Equity Premium of 6.00%

Financial Economics, Time Variation in the Market Return, Figure 2
Bivariate scatterplots of simulated data for a model allowing for trends and structural breaks. The model upon which these scatterplots are based allows for trends and structural breaks in the equity premium process, as well as autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. Ex ante equity premium of 2.5%: ⋄, Ex ante equity premium of 3.5%: ∘, Ex ante equity premium of 4.5%: □, Ex ante equity premium of 6%: ⊕

simulated pairs in each panel is surrounded by an ellipse which represents a 95% bivariate confidence bound, based on the asymptotic normality (or log-normality, where appropriate) of the plotted variables. (The 95% confidence ellipsoids are asymptotic approximations based on joint normality of the sample estimates of the moments of the simulated data. All of the sample moment estimates we consider are asymptotically normally distributed, as can be seen by appealing to the appropriate law of large numbers). The confidence ellipse for the 2.5% case is marked with diamonds, the 3.5% case with circles, the 4.5% case with squares, and the 6% case with circled crosses.

Notice that the sample mean for the US economy (the intersection of the crosshairs) lies loosely within cloud of points that depict the set of simulated economies for each ex ante equity premium case. That is, our simulations produce mean returns and return volatility that roughly match the US observed moments of returns, *without our having calibrated to returns*. Notice also that the intersection of the crosshairs is outside (or very nearly outside) the 95% confidence ellipse in all cases except that of the 3.5% ex ante equity premium. (In unreported results that study a finer grid of ex ante equity premium values, we found that only those simulations based on values of the ex ante equity premium between about 2.5% and 4.5% lead to 95% confidence ellipses that encompass the US economy crosshairs. As the value of the ex ante equity premium falls below 2.5% or rises above 4.5%, the confidence ellipse drifts further away from the crosshairs). Based on this set of plots, we can conclude that ex ante equity premia much less than or much greater than 3.5% are inconsistent at the 5% confidence level with the observed mean return and return volatility of S&P 500 returns. χ^2 tests presented in Donaldson, Kamstra, and Kramer [16] confirm this result.

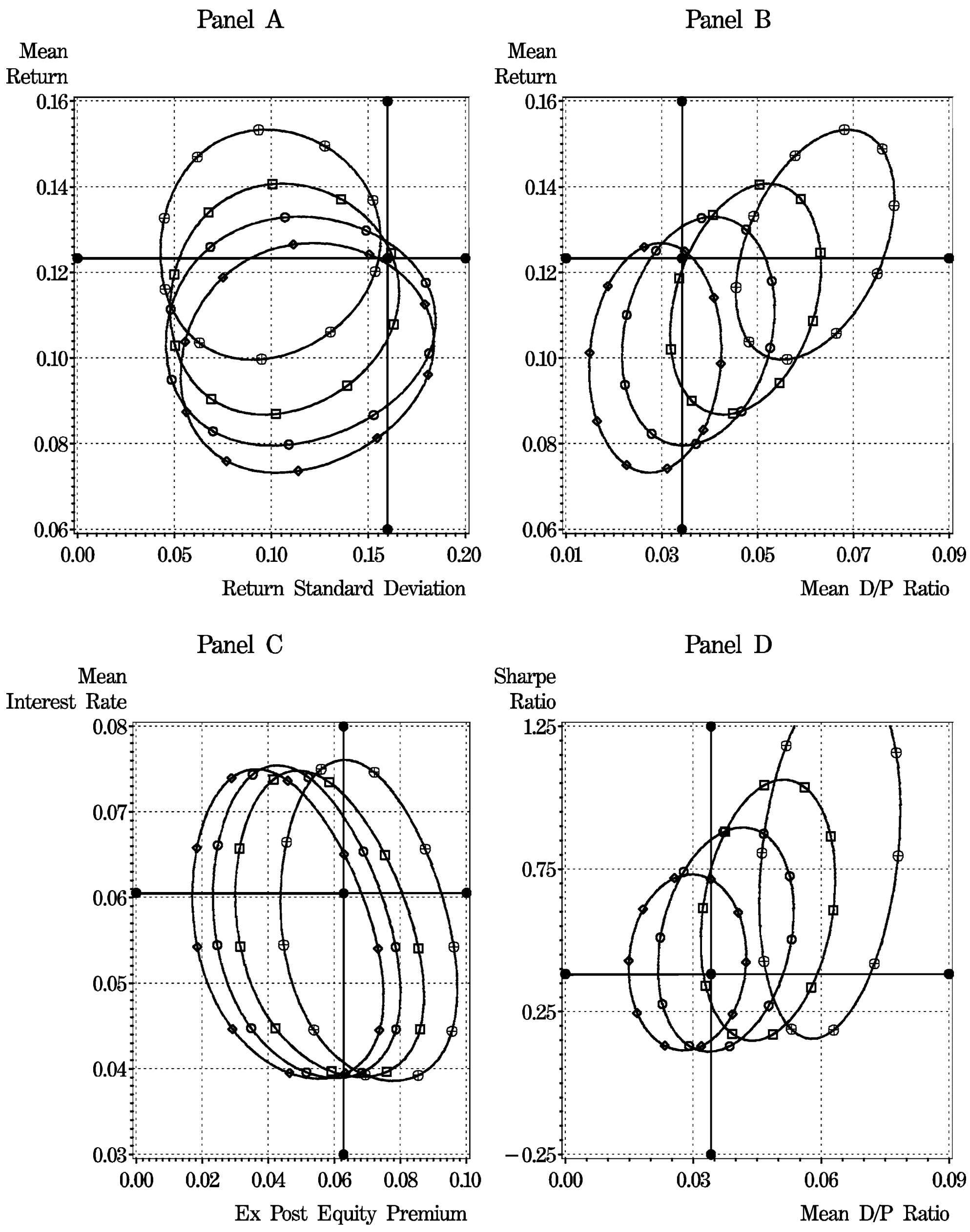
We can easily condense the information contained in these four individual plots into one plot, as shown in Panel A of Fig. 3. The scatterplot of points representing individual simulations are omitted in the condensed plot, but the confidence ellipses themselves (and the symbols used to distinguish between them) are retained. Panel A of Fig. 3 repeats the ellipses shown in Fig. 2, so that again we see that only the 3.5% ex ante equity premium case is well within the confidence ellipse at the 5% significance level. In presenting results for additional bivariate combinations, we follow the same practice, omitting the points that represent individual simulations and using the same set of symbols to distinguish between confidence ellipses based on ex ante equity premia of 2.5%, 3.5%, 4.5%, and 6%.

In Panel B of Fig. 3 we consider the four sets of confidence ellipses for mean return and mean dividend yield combinations. Notice that as we increase the ex ante equity premium, the confidence ellipses shift upward and to the right. Notice also that with higher values of the ex ante equity premium we tend to have more variable dividend yields. That is, the confidence ellipse covers a larger range of dividend yields when the value of the ex ante equity premium is larger. The observed combination of S&P 500 mean return and mean dividend yield, represented by the intersecting crosshairs, lies within the confidence ellipse for the 2.5% and 3.5% cases, very close to the ellipse for the 4.5% case, and far outside the ellipse for the 6% case.

Panel C of Fig. 3 plots confidence ellipses for mean interest rates versus mean ex post equity premia. The intersection of the crosshairs is within all four of the shown confidence ellipses. As we calibrate our model to the US interest rate, and as the ex post equity premium has a large variance, it is not surprising that the US experience is consistent with the simulated data from the entire range of ex ante equity premia considered here. This result is merely telling us that the ex post equity premium is not, by itself, particularly helpful in narrowing the possible range for the ex ante equity premium (consistent with the empirical imprecision in measuring the ex post equity premium which has been extensively documented in the literature). Notice as well that the confidence ellipses in Panel C are all negatively sloped: we see high mean interest rates with low equity premia and low mean interest rates with high equity premia. Many researchers, including Weil [74], have commented that the flip side of the high equity premium puzzle is the low risk-free rate puzzle. Here we confirm that the dual puzzle arises in our simulated economies as well. It appears that this puzzle is a mechanical artifact coming out of the calculation of the premium. As the ex post equity premium equals the mean return minus the mean interest rate, a decrease in the interest rate, all else held constant, must lead to a higher ex post equity equity premium.

Panel D of Fig. 3 contains the confidence ellipses for the Sharpe ratio (or reward-to-risk ratio, calculated as the average annual difference between the arithmetic return and the risk-free rate divided by the standard deviation of the annual differences) and the mean dividend yield. As the ex ante equity premium is increased from 2.5%, the confidence ellipses shift from being centered on the crosshairs to far to the right of the crosshairs. The US experience, indicated by the crosshairs at a Sharpe ratio of approximately 0.4 and a mean dividend yield of about 3.5%, is well outside the 95% confidence ellipse for the 6% ex ante equity premium case, suggesting a 6% ex ante equity premium is inconsistent with the jointly observed S&P 500 Sharpe ratio and mean dividend yield. Indeed Fama and French [21] and Jagannathan, McGrattan, and Scherbina [35] make reference to dividend yields to argue that the equity premium may be much smaller than 6%; our analysis gives us a glimpse of just how much smaller it might be.

Overall in Fig. 3, the joint realization of key characteristics of the US market data suggests that the true ex ante equity premium is no lower than 2.5%, no higher than 4.5%, and is most likely near 3.5%. Multivariate χ^2 tests performed by Donaldson, Kamstra, and Kramer [16] indi-

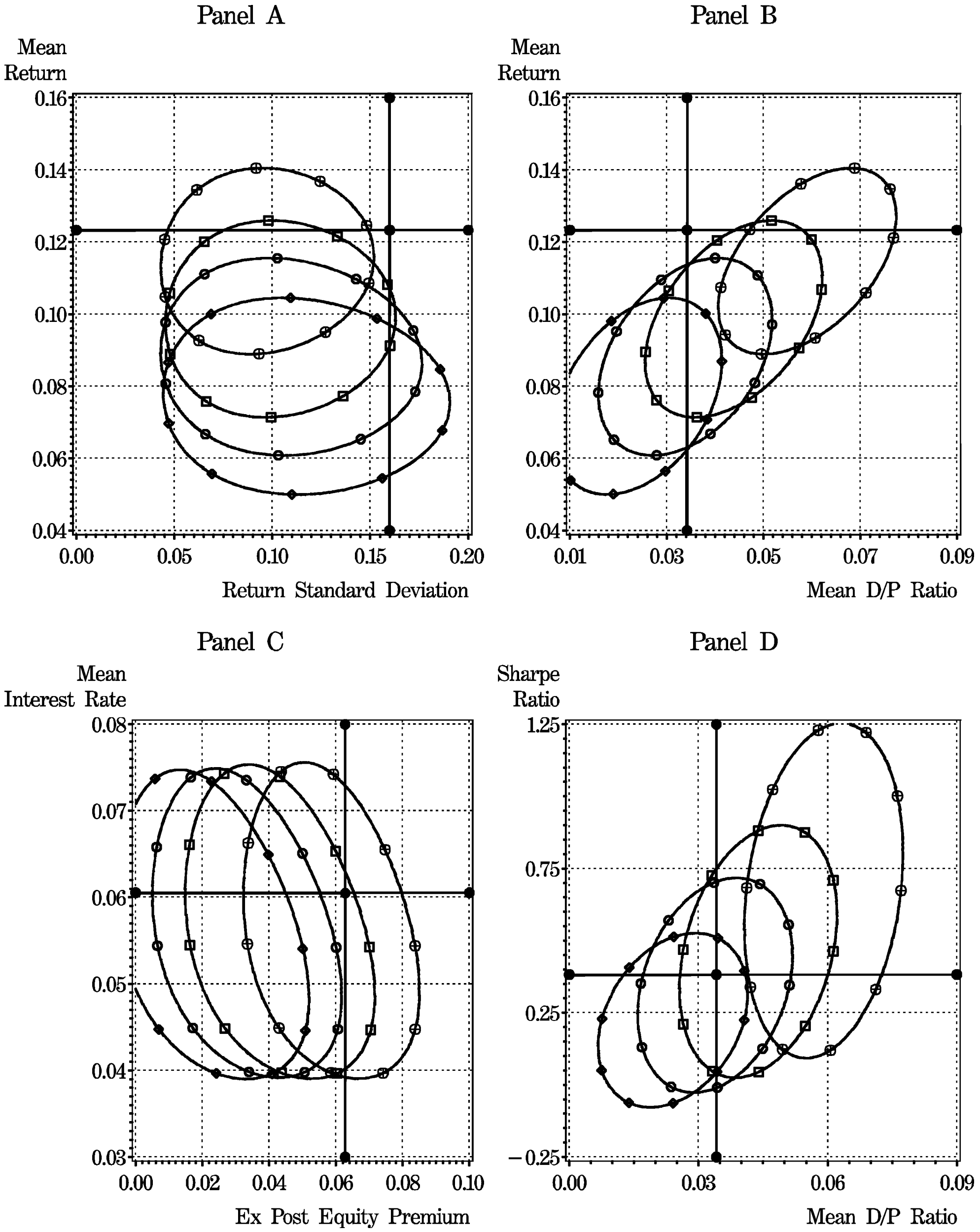


Financial Economics, Time Variation in the Market Return, Figure 3

Confidence ellipses based on simulated data for a model allowing for trends and structural breaks. The model upon which these scatterplots are based allows for trends and structural breaks in the equity premium process, as well as autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. 2.5% ex post equity premium: ⋄, 3.5% ex post equity premium: ○, 4.5% ex post equity premium: □, 6% ex post equity premium: ⊕

cate a 95% confidence interval of plus-or-minus 50 basis points around 3.5%.

Consider now Fig. 4, which details simulated data from a restricted model that has a time-varying equity premium but no trends or structural breaks. Donaldson, Kamstra, and Kramer [16] study this simplified model and find that it performs poorly relative to the model we consider in Figs. 2 and 3 in terms of its ability to capture the behavior of US market data. Figure 4 shows that with the restricted model, no values of the ex ante equity premium are



Financial Economics, Time Variation in the Market Return, Figure 4

Confidence ellipses based on simulated data for a restricted model that does not allow for trends and structural breaks. The model upon which these scatterplots are based does not allow for trends or structural breaks in the equity premium process, but does allow for autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. 2.5% ex post equity premium: ⋄, 3.5% ex post equity premium: ○, 4.5% ex post equity premium: □, 6% ex post equity premium: ⊕

consistent with the observed US mean return, standard deviation, and dividend yield. That is, the simulation-based mean return and dividend yield ellipses do not contain the US data crosshairs for any value of the ex ante equity premium considered. (χ^2 tests presented in Donaldson, Kamstra, and Kramer [16] strongly support this conclusion). The implication is that it is essential to model trends and structural breaks in the equity premium process in order to accurately capture the dynamics of observed US data. Donaldson, Kamstra, and Kramer show that model failure becomes even more stark if the equity premium is constrained to be constant.

Overall, the evidence in Figs. 3 and 4 does not itself resolve the equity premium puzzle, but evidence in Fig. 3 (based on the model that allows for trends and structural breaks in the equity premium process) does provide a narrow target range of plausible equity premia that economic models should be able to explain. Additionally, the evidence in Figs. 3 and 4 points to a secondary issue ignored in the literature prior to the work of Donaldson, Kamstra, and Kramer [16], that it is crucial to model the equity premium as both time-varying and as having trends and structural breaks. We saw in Fig. 4 that high return volatility, high ex post equity premia, and low dividend yields cannot be explained easily by constant equity premium models. This result has clear implications for valuation: simple techniques that restrict the discount rate to a constant are remarkably inconsistent with the US experience of time-varying equity premia, and serious attention should be paid to modeling a time-varying rate for use in discounting future expected cash flows.

Time-Varying Equity Premia: Possible Biological Origins

To the extent that the simulation techniques considered in the previous section suggest that the equity premium varies over time, it is interesting to consider some empirical evidence of time-varying equity premia. We first survey some examples of high-frequency variations in the equity premium, and then we explore in detail two examples which may arise due to reasons that relate to human biology and/or psychology.

There is a wide range of evidence of high-frequency movement in the equity premium. At the highest frequency, we observe roughly 'U-shaped' intra-day returns (see [29,36,77]), with returns being perhaps somewhat higher during the morning trading period than in the afternoon (see [46]). At the weekly frequency, returns from Friday's close until Monday's close are low and even negative on average, as first identified by Cross [12]. Rogalski [66] found prices rose during Mondays, thus identifying the negative average realizations that followed Fridays as a weekend effect and not a Monday effect. Turning to the monthly domain, Ogden [56] documented a turn of the month effect where returns in the first half of the month are higher than returns in the second half of the month. At the annual frequency, there is the well-known turn-of-the-year effect, first shown by Rozeff and Kinney [68]. Keim [45] showed that half of the year's excess returns for small firms arose in January, and half of the January returns took place in the first five days of the month. Further, Reinganum [64] showed that January returns are higher for small firms whose price performed poorly in the previous year. All of this is consistent with the tax-loss-selling hypothesis whereby investors realize losses at the end of the tax year, leading to higher returns in January after the tax-loss selling ends.

Next we turn our attention to two cases of time-varying equity premia that may arise for reasons related to human physiology. One is Seasonal Affective Disorder (SAD), and another is daylight saving time changes.

Seasonal Affective Disorder

Past research suggests there are seasonal patterns in the equity premium which may arise due to cyclical changes in the risk tolerance of individual investors over the course of the year related to SAD. The medical condition of SAD, according to Rosenthal [67], is a recurrent depression associated with diminished daylight in the fall, affecting many millions of Americans, as well as peoples from around the world, even those located near the equator. (In a study of 303 patients attending a primary care facility in Vancouver, Schlager, Froom, and Jaffe [70] found that 9% were clinically diagnosed with SAD and another 29% had significant winter depressive symptoms without meeting conditions for major depression. Other studies have found similar magnitudes, though some research has found that prevalence varies by latitude, with more extreme latitudes having a larger proportion of SAD-sufferers.) SAD is classified as a major depressive disorder. The symptoms of SAD include anxiety, periods of sadness, chronic fatigue, difficulty concentrating, lethargy, sleep disturbance, sugar and carbohydrate craving and associated weight gain, loss of interest in sex, and of course, clinical depression. Psychologists have shown that depressed people have less tolerance for risk in general. (See [7,32,82,83]). Psychologists refer to risk tolerance in terms of "sensation seeking" tendency, measured using a scale developed by Zuckerman [80], [81]. Those who tolerate (or seek) high levels of risk tend to score high on the sensation-seeking scale. Differences in sensation-seeking tendencies have been linked to gender (see [5] for example), race (see [31] for instance), age (see, for example, [84]), and other personal characteristics.

Economists and psychologists working together have shown that sensation-seeking tendency translates into tolerance for risk of a specifically financial or economic nature. For instance, Wong and Carducci [76] find that individuals who score low on tests of sensation seeking display greater risk aversion in making financial decisions, including the decision to purchase stocks, bonds, and insurance. Harlow and Brown [28] document the link between sensation seeking and financial risk tolerance by building on results from psychiatry which show that high blood levels of a particular enzyme are associated with depression and a lack of sensation seeking while low levels of the enzyme are associated with a high degree of sensation seeking. Harlow and Brown write, "Individuals with neurochemical activity characterized by lower levels of [the enzyme] and with a higher degree of sensation-seeking are *more willing to accept economic risk*… Conversely, high levels of this enzyme and a low level of sensation seeking appear to be associated with risk-averse behavior." (pp. 50–51, emphasis added). These findings suggest an individual's level of sensation seeking is indicative of his or her tolerance for financial risk.

Given these relationships, Kamstra, Kramer, and Levi [42] conjecture that during the fall and winter seasons, when a fraction of the population suffers from SAD, the proportion of risk-averse investors rises. Risk-averse investors shun risky stocks in the fall, they argue, which has a negative influence on stock prices and returns. As winter progresses and daylight becomes more plentiful, investors start to recover from their depression and become more willing to hold risky assets, at which time stock prices and returns should be positively influenced.

If the extent or severity of SAD is greater at more extreme latitudes, then the SAD effect on stock returns should be greater in stock markets at high latitudes and less in markets close to the equator. Also, the pattern of returns in the Southern Hemisphere should be the opposite of that in the Northern Hemisphere as are the seasons. Thus, Kamstra, Kramer and Levi [42] study stock market indices for the US, Sweden, Britain, Germany, Canada, New Zealand, Japan, Australia, and South Africa. They regress each country's daily stock returns on a variety of standard control variables plus two variables intended to capture the impact of SAD on returns. The first of these two variables, SAD_t, is a simple function of the length of night at the latitude of the respective market for the fall and winter months for which SAD has been documented to be most severe. The second of these variables, a fall dummy variable denoted Fall_t, is included because the SAD hypothesis implies the expected effect on returns is different before versus after winter solstice. Specifically, when agents initially become more risk averse, they should shun risky assets which should cause prices to be lower than would otherwise be observed, and when agents revert to normal as daylight becomes more plentiful, prices should rebound. The result should be lower returns in the autumn, higher returns in the winter, and thus a high equity premium for investors who hold through the autumn and winter periods. The Fall_t dummy variable is used to capture the lower autumn returns. Both SAD_t and Fall_t are appropriately defined for the Southern Hemisphere countries, accounting for the six month difference in seasons relative to the Northern Hemisphere markets.

Table 1 summarizes the average annual effect due to each of the SAD variables, SAD_t and Fall_t, for each of the international indices Kamstra, Kramer, and Levi [42] study. For comparison, the unconditional average annual return for each index is also provided. Observe that the annualized return due to SAD_t is positive in every country, varying from 5.7 to 17.5 percent. The SAD effect is generally larger the further are the markets from the equator. The negative annualized returns due to Fall_t demonstrate the fact that SAD typically causes returns to be shifted from the fall to the winter. Garrett, Kamstra, and Kramer [24] study seasonally-varying risk aversion in the context of an equilibrium asset pricing model, allowing the price of risk to vary with length of night through the fall and winter seasons. They find the risk premium on equity varies through the seasons in a manner consistent with investors being more risk averse due to SAD in the fall and winter.

Kamstra, Kramer, and Levi [43] show that there is an opposite seasonal pattern in Treasury returns relative to stock returns, consistent with time-varying risk aversion being the underlying force behind the seasonal pattern previously shown to exist in stock returns. If SAD-affected investors are shunning risky stocks in the fall as they become more risk averse, then they should be favoring safe assets at that time, which should lead to an opposite pattern in Treasury returns relative to stock returns. The seasonal cycle in the Treasury market is striking, with a variation of more than 80 basis points between the highest and lowest average monthly returns. The highest Treasury returns are observed when equity returns are lowest, and *vice versa*, which is a previously unknown pattern in Treasury returns.

Kamstra, Kramer, and Levi [43] define a new measure which is linked directly to the clinical incidence of SAD. The new measure uses data on the weekly or monthly onset of and recovery from SAD, obtained from studies of SAD patients in Vancouver and Chicago conducted by medical researchers. Young, Meaden, Fogg, Cherin, and Eastman [79] and Lam [47] document the clinical *onset* of SAD symptoms and *recovery* from SAD symptoms among North Americans known to be affected by SAD. Young et al. study 190 SAD-sufferers in Chicago and find that 74 percent of them are first diagnosed with SAD in the weeks between mid-September and early November. Lam

Financial Economics, Time Variation in the Market Return, Table 1
Average annual percentage return due to SAD variables

Country (Index)	Annual return due to SAD_t	Annual return due to fall_t	Unconditional annual return
US (S&P 500)	9.2^{***}	-3.6^{**}	6.3^{***}
Sweden (Veckans Affärar)	13.5^{**}	-6.9^{**}	17.1^{***}
Britain (FTSE 100)	10.3^{**}	-2.3	9.6^{***}
Germany (DAX 30)	8.2^{*}	-4.3^{**}	6.5^{**}
Canada (TSX 300)	13.2^{***}	-4.3^{**}	6.1^{***}
New Zealand (Capital 40)	10.5^{**}	-6.6^{**}	3.3
Japan (NIKKEI 225)	6.9^{*}	-3.7^{**}	9.7^{***}
Australia (All ordinaries)	5.7	0.5	8.8^{***}
South Africa (Datastream global index)	17.5^{*}	-2.1	14.6^{***}

One, two, and three asterisks denote significantly different from zero at the ten, five, and one percent level respectively, based on one-sided tests. Source: Table 3 in [42].

studies 454 SAD patients in Vancouver on a monthly basis and finds, that the peak timing of diagnosis is during the early fall. Lam [47] also studies the timing of clinical remission of SAD and finds it peaks in April, with almost half of all SAD-sufferers first experiencing complete remission in that month. March is the second most common month for subjects to first experience full remission, corresponding to almost 30 percent of subjects. For most SAD patients, the initial onset and full recovery are separated by several months over the fall and winter.

Direct use of Kamstra, Kramer, and Levi's [43] variable (which is an estimate of population-wide SAD onset/recovery based on specific samples of individuals) could impart an error-in-variables problem (see [48]), thus they utilize an instrumented version detailed in the paper, which they call Onset/Recovery, denoted $\hat{O}R_t$. The instrumented SAD measure $\hat{O}R_t$ reflects the change in the proportion of SAD-affected individuals actively suffering from SAD. The measure is defined year-round (unlike the original Kamstra, Kramer, and Levi [42], SAD_t variable, which is defined for only the fall and winter months), taking on positive values in the summer and fall and negative values in the winter and spring. Its value peaks near the fall equinox and reaches a trough near the spring equinox. (The exact monthly values of $\hat{O}R_t$ are reported by Kamstra, Kramer, and Levi [43].) The opposite signs on $\hat{O}R_t$ across the fall and winter seasons should, in principle, permit it to capture the opposite impact on equity or Treasury returns across the seasons, without use of a dummy variable. Kamstra, Kramer, and Levi [43] find that use of $\hat{O}R_t$ as a regressor to explain seasonal patterns in Treasury and equity returns renders the SAD_t and Fall_t (used by Kamstra, Kramer, and Levi [42]) as economically and statistically insignificant, suggesting the Onset/Recovery variable does a far better job of explaining seasonal variation in returns than the original proxies which are not directly related to the incidence of SAD.

Kamstra, Kramer, and Levi [43] show that the seasonal Treasury and equity return patterns are unlikely to arise from macroeconomic seasonalities, seasonal variation in risk, cross-hedging between equity and Treasury markets, investor sentiment, seasonalities in the Treasury market auction schedule, seasonalities in the Treasury debt supply, seasonalities in the Federal Reserve Board's interest-rate-setting cycle, or peculiarities of the sample period considered. They find that the seasonal cycles in equity and Treasury returns become more pronounced during periods of high market volatility, consistent with time-varying risk aversion among market participants. Furthermore, they apply the White [75] reality test and find that the correlation between returns and the clinical incidence of seasonal depression cannot be easily dismissed as the simple result of data snooping.

DeGennaro, Kamstra, and Kramer [13] and Kamstra, Kramer, and Levi [13] provide further corroborating evidence for the hypothesis that SAD leads to time variation in financial markets by considering (respectively) bid-ask spreads for stocks and the flow of funds in and out of risky and safe mutual funds. In both papers they find strong support for the link between seasonal depression and time-varying risk aversion.

Daylight Saving Time Changes

The second potential biological source of time-varying equity premia we consider arises on the two dates of the year when most of the developed world shifts clocks forward or backward an hour in the name of daylight saving. Psychologists have found that changes in sleep patterns (due to shift work, jet lag, or daylight saving time changes, for example) are associated with increased anxiety, which is suggestive of a link between changes in sleep habits and time-varying risk tolerance. See [26,52], and citations found in [10] and [72] for more details on the link between sleep disruptions and anxiety. In addition to causing heightened anxiety, changes in sleep patterns also inhibit rational decision-making, lower one's information-processing ability, affect judgment, slow reaction time, and reduce problem-solving capabilities. Even a change of one hour can significantly affect behavior.

Kamstra, Kramer, and Levi [40] explore the financial market ramifications of a link between daylight saving time-change-induced disruptions in sleep patterns and individuals' tolerance for risk. They find, consistent with psychology studies that show a gain or loss of an hour's sleep leads to increased anxiety, investors seem to shun risky stock on the trading day following a daylight saving time change. They consider stock market indexes from four countries where the time changes happen on non-overlapping dates, the US, Canada, Britain, and Germany. Based on stock market behavior over the past three decades, the authors find that the magnitude of the average return on spring daylight saving weekends is typically between two to five times that of ordinary weekends, and the effect is even stronger in the fall. Kamstra, Kramer, and Levi [41] show that the effect is not driven by a few extremely negative observations, but rather the entire distribution of returns shifts to the left following daylight saving time changes, consistent with anxious investors selling risky stock.

Future Directions

We divide our discussion in this section into three parts, one for each major topic discussed in the article.

Regarding fundamental valuation, a promising future path is to compare estimates emerging from sophisticated valuation methods to market prices, using the comparison to highlight inconsistencies in the modeling assumptions (such as restrictions on the equity premium used by the model, restrictions on the growth rate imposed for expected cash flows, and the implied values of those quantities that can be inferred from market prices). Even if one believes that markets are efficient and investors are rational, there is still much to be learned from calculating fundamentals using models and examining discrepancies relative to observed market prices.

Regarding the simulation techniques for estimating the equity premium, a promising direction for future research is to exploit these tools to forecast the volatility of stock prices. This may lead to new alternatives to existing option-implied volatility calculations and time-series techniques such as ARCH (for an overview of these methods see [15]). Another fruitful future direction would be to apply the simulation techniques to the valuation of individual companies' stock (as opposed to valuing, say, stock market indexes).

Regarding the topic of time-varying equity premia that may arise for biological reasons, a common feature of both of the examples explored in Sect. "Time-Varying Equity Premia: Possible Biological Origins", SAD and daylight-saving-time-change-induced fluctuations in the risk premium, is that in both cases the empirical evidence is based on aggregate financial market data. There is a recent trend in finance toward documenting phenomena at the individual level, using data such as individuals' financial asset holdings and trades in their brokerage accounts. (See [1,54,55] for instance). A natural course forward is to build upon the existing aggregate market support for the prevalence of time-varying risk aversion by testing at the individual level whether risk aversion varies through the course of the year due to seasonal depression and during shorter intervals due to changes in sleep patterns. An additional potentially fruitful direction for future research is to integrate into classical asset pricing models the notion that biological factors might impact asset returns through changes in agents' degree of risk aversion. That is, human traits such as seasonal depression may lead to regularities in financial markets that are not mere anomalies; rather they may be perfectly consistent with rational agents making sensible decisions given their changing tolerance for risk. This new line of research would be similar in spirit to the work of Shefrin [71] who considers the way behavioral biases like overconfidence can be incorporated into the pricing kernel in standard asset pricing models. While the behavioral biases Shefrin considers typically involve humans making errors, the biological factors described here might be considered rational due to their involvement of time-varying risk aversion.

Bibliography

Primary Literature

1. Barber B, Odean T (2001) Boys will be boys: Gender, overconfidence, and common stock investment. Q J Econ 116: 261–292
2. Barsky RB, DeLong JB (1993) Why does the stock market fluctuate? Q J Econ 108:291–311
3. Bakshi G, Chen Z (2005) Stock valuation in dynamic economies. J Financ Market 8:115–151

4. Brealey RA, Myers SC, Allen F (2006) Principles of corporate finance, 8th edn. McGraw-Hill Irwin, New York
5. Byrnes JP, Miller DC, Schafer WD (1999) Gender differences in risk taking: A meta-analysis. Psychol Bull 125:367–383
6. Campbell JY, Kyle AS (1993) Smart money, noise trading and stock price behavior. Rev Econ Stud 60:1–34
7. Carton S, Jouvent R, Bungener C, Widlöcher D (1992) Sensation seeking and depressive mood. Pers Individ Differ 13: 843–849
8. Chiang R, Davidson I, Okuney J (1997) Some theoretical and empirical implications regarding the relationship between earnings, dividends and stock prices. J Bank Financ 21:17–35
9. Cochrane JH (2001) Asset pricing. Princeton University Press, Princeton
10. Coren S (1996) Sleep Thieves. Free Press, New York
11. Corradi V, Swanson NR (2005) Bootstrap specification tests for diffusion processes. J Econom 124:117–148
12. Cross F (1973) The behavior of stock prices on Fridays and Mondays. Financ Anal J 29:67–69
13. DeGennaro R, Kamstra MJ, Kramer LK (2005) Seasonal variation in bid-ask spreads. University of Toronto (unpublished manuscript)
14. Donaldson RG, Kamstra MJ (1996) A new dividend forecasting procedure that rejects bubbles in asset prices. Rev Financ Stud 9:333–383
15. Donaldson RG, Kamstra MJ (2005) Volatility forecasts, trading volume, and the ARCH versus option-implied volatility tradeoff. J Financ Res 28:519–538
16. Donaldson RG, Kamstra MJ, Kramer LA (2007) Estimating the ex ante equity premium. University of Toronto Manuscript
17. Dong M, Hirshleifer DA (2005) A generalized earnings-based stock valuation model. Manchester School 73:1–31
18. Duffie D, Singleton KJ (1993) Simulated moments estimation of Markov models of asset prices. Econometrica 61:929–952
19. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
20. Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J Financ Econ 33:3–56
21. Fama EF, French KR (2002) The equity premium. J Financ 57:637–659
22. Farrell JL (1985) The dividend discount model: A primer. Financ Anal J 41:16–25
23. Feltham GA, Ohlson JA (1995) Valuation and clean surplus accounting for operating and financial activities. Contemp Account Res 11:689–731
24. Garrett I, Kamstra MJ, Kramer LK (2005) Winter blues and time variation in the price of risk. J Empir Financ 12:291–316
25. Gordon M (1962) The Investment, Financing and Valuation of the Corporation. Irwin, Homewood
26. Gordon NP, Cleary PD, Parker CE, Czeisler CA (1986) The prevalence and health impact of shiftwork. Am J Public Health 76:1225–1228
27. Hackel KS, Livnat J (1996) Cash flow and security analysis. Irwin, Chicago
28. Harlow WV, Brown KC (1990) Understanding and assessing financial risk tolerance: A biological perspective. Financ Anal J 6:50–80
29. Harris L (1986) A transaction data study of weekly and intradaily patterns in stock returns. J Financ Econ 16:99–117
30. Hawkins DF (1977) Toward an old theory of equity valuation. Financ Anal J 33:48–53
31. Hersch J (1996) Smoking, seat belts and other risky consumer decisions: differences by gender and race. Managerial Decis Econ 17:471–481
32. Horvath P, Zuckerman M (1993) Sensation seeking, risk appraisal, and risky behavior. Personal Individ Diff 14:41–52
33. Hurley WJ, Johnson LD (1994) A realistic dividend valuation model. Financ Anal J 50:50–54
34. Hurley WJ, Johnson LD (1998) Generalized Markov dividend discount models. J Portf Manag 24:27–31
35. Jagannathan R, McGrattan ER, Scherbina A (2000) The declining US equity premium. Fed Reserve Bank Minneap Q Rev 24:3–19
36. Jain PC, Joh G (1988) The dependence between hourly prices and trading volume. J Financ Quant Analysis 23:269–283
37. Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. J Financ 48:65–91
38. Jorion P, Goetzmann WN (1999) Global stock markets in the twentieth century. J Financ 54:953–980
39. Kamstra MJ (2003) Pricing firms on the basis of fundamentals. Fed Reserve Bank Atlanta Econ Rev First Quarter:49–70
40. Kamstra MJ, Kramer LA, Levi MD (2000) Losing sleep at the market: the daylight saving anomaly. Am Econ Rev 90: 1005–1011
41. Kamstra MJ, Kramer LA, Levi MD (2002) Losing sleep at the market: The daylight saving anomaly: Reply. Am Econ Rev 92:1257–1263
42. Kamstra MJ, Kramer LA, Levi MD (2003) Winter blues: A SAD stock market cycle. Am Econ Rev 93:324–343
43. Kamstra MJ, Kramer LA, Levi MD (2007) Opposing Seasonalities in Treasury versus Equity Returns. University of Toronto Manuscript
44. Kamstra MJ, Kramer LA, Levi MD, Wermers R (2008) Seasonal asset allocation: evidence from mutual fund flows. University of Toronto Manuscript
45. Keim DB (1983) Size-related anomalies and stock return seasonality: Further Empirical Evidence. J Financ Econ 12:13–32
46. Kramer LA (2002) Intraday stock returns, time-varying risk premia, and diurnal mood variation. University of Toronto Manuscript
47. Lam RW (1998) Seasonal Affective Disorder: Diagnosis and management. Prim Care Psychiatry 4:63–74
48. Levi MD (1973) Errors in the variables bias in the presence of correctly measured variables. Econometrica 41:985–986
49. Liu W (2006) A liquidity-augmented capital asset pricing model. J Financ Econ 82:631–671
50. McFadden D (1989) A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica 47:995–1026
51. Mehra R, Prescott EC (1985) The equity premium: A puzzle. J Monet Econ 15:145–161
52. Mellinger GD, Balter MB, Uhlenhuth EH (1985) Insomnia and its treatment: prevalence and correlates. Arch Gen Psychiatry 42:225–232
53. Michaud RO, Davis PL (1982) Valuation model bias and the scale structure of dividend discount returns. J Financ 37: 563–576
54. Odean T (1998) Are investors reluctant to realize their losses? J Financ 53:1775–1798

55. Odean T (1999) Do investors trade too much? Am Econ Rev 89:1279–1298
56. Ogden JP (1990) Turn-of-month evaluations of liquid profits and stock returns: A common explanation for the monthly and January effects. J Financ 45:1259–1272
57. Ohlson JA (1995) Earnings, book values, and dividends in equity valuation. Contemp Acc Res 11:661–687
58. Pakes A, Pollard D (1989) Simulation and the asymptotics of optimization estimators. Econometrica 57:1027–1057
59. Pástor L', Stambaugh R (2001) The equity premium and structural breaks. J Financ 56:1207–1239
60. Penman SH (1998) Combining earnings and book value in equity valuation. Contemp Acc Res 15:291–324
61. Penman SH, Sougiannis T (1998) A comparison of dividend, cash flow and earnings approaches to equity valuation. Contemp Acc Res 15:343–383
62. Peters DJ (1991) Using PE/Growth ratios to develop a contrarian approach to growth stocks. J Portf Manag 17:49–51
63. Rappaport A (1986) The affordable dividend approach to equity valuation. Financ Anal J 42:52–58
64. Reinganum MR (1983) The anomalous stock market behavior of small firms in January. J Financ Econ 12:89–104
65. Rietz TA (1988) The equity risk premium: A solution. J Monet Econ 22:117–31
66. Rogalski RJ (1984) New findings regarding day-of-the-week returns: A note. J Financ 35:1603–1614
67. Rosenthal NE (1998) Winter Blues: Seasonal Affective Disorder: What is It and How to Overcome It, 2nd edn. Guilford Press, New York
68. Rozeff MS, Kinney WR (1976) Capital market seasonality: The case of stock returns. J Financ Econ 3:379–402
69. Rubinstein M (1976) The valuation of uncertain income streams and the pricing of options. Bell J Econ 7:407–425
70. Schlager D, Froom J, Jaffe A (1995) Winter depression and functional impairment among ambulatory primary care patients. Compr Psychiatry 36:18–24
71. Shefrin H (2005) A Behavioral Approach to Asset Pricing. Academic Press, Oxford
72. Spira AP, Friedman L, Flint A, Sheikh J (2005) Interaction of sleep disturbances and anxiety in later life: perspectives and recommendations for future research. J Geriatr Psychiatry Neurol 18:109–115
73. Sorensen EH, Williamson DA (1985) Some evidence on the value of dividend discount models. Financ Anal J 41:60–69
74. Weil P (1989) The equity premium puzzle and the risk-free rate puzzle. J Monet Econ 24:401–421
75. White H (2000) A reality check for data snooping. Econometrica 68:1097–1126
76. Wong A, Carducci B (1991) Sensation seeking and financial risk taking in everyday money matters. J Bus Psychol 5: 525–530
77. Wood RA, McInish TH, Ord JK (1985) An investigation of transactions data for NYSE stocks. J Financ 40:723–741
78. Yao Y (1997) A trinomial dividend valuation model. J Portf Manag 23:99–103
79. Young MA, Meaden PM, Fogg LF, Cherin EA, Eastman CI (1997) Which environmental variables are related to the onset of Seasonal Affective Disorder? J Abnorm Psychol 106: 554–562
80. Zuckerman M (1976) Sensation seeking and anxiety, traits and states, as determinants of behavior in novel situations. In: Sarason IG, Spielberger CD (eds) Stress and Anxiety, vol 3. Hemisphere, Washington DC
81. Zuckerman M (1983) Biological Bases of Sensation Seeking, Impulsivity and Anxiety. Lawrence Erlbaum Associates, Hillsdale
82. Zuckerman M (1984) Sensation seeking: A comparative approach to a human trait. Behav Brain Sci 7:413–471
83. Zuckerman M, Buchsbaum MS, Murphy DL (1980) Sensation seeking and its biological correlates. Psychol Bull 88: 187–214
84. Zuckerman M, Eysenck S, Eysenck HJ (1978) Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. J Consult Clin Psychol 46:139–149

Books and Reviews

Dimson E (1988) Stock Market Anomalies. Cambridge University Press, Cambridge

Kocherlakota NR (1996) The equity premium: It's still a puzzle. J Econ Lit 34:42–71

Mehra R (2003) The equity premium: Why is it a puzzle? Financ Anal J 59:54–69

Mehra R, Prescott EC (2003) The equity premium in retrospect. In: Constantinides GM, Harris M, Stulz RM (eds) Handbook of the Economics of Finance: Financial Markets and Asset Pricing, vol 1B. North Holland, Amsterdam, pp 889–938

Penman S (2003) Financial Statement Analysis and Security Valuation, 2nd edn. McGraw-Hill/Irwin, New York

Siegel JJ, Thaler RH (1997) Anomalies: The equity premium puzzle. J Econ Perspect 11:191–200

Thaler RH (2003) The Winner's Curse: Paradoxes and Anomalies of Economic Life. Princeton University Press, Princeton

Financial Forecasting, Non-linear Time Series in

Gloria González-Rivera, Tae-Hwy Lee
Department of Economics, University of California, Riverside, USA

Article Outline

Glossary

Arbitrage pricing theory (APT) the expected return of an asset is a linear function of a set of factors.

Artificial neural network is a nonlinear flexible functional form, connecting inputs to outputs, being capable of approximating a measurable function to any desired level of accuracy provided that sufficient complexity (in terms of number of hidden units) is permitted.

Autoregressive conditional heteroskedasticity (ARCH) the variance of an asset returns is a linear function of the past squared surprises to the asset.

Bagging short for *b*ootstrap *agg*regat*ing*. Bagging is a method of smoothing the predictors' instability by averaging the predictors over bootstrap predictors and thus lowering the sensitivity of the predictors to training samples. A predictor is said to be unstable if perturbing the training sample can cause significant changes in the predictor.

Capital asset pricing model (CAPM) the expected return of an asset is a linear function of the covariance of the asset return with the return of the market portfolio.

Factor model a linear factor model summarizes the dimension of a large system of variables by a set of factors that are linear combinations of the original variables.

Financial forecasting prediction of prices, returns, direction, density or any other characteristic of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc.

Functional coefficient model a model with time-varying and state-dependent coefficients. The number of states can be infinite.

Linearity in mean the process $\{y_t\}$ is linear in mean conditional on X_t if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k .$$

Loss (cost) function When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c_{t+h}(Y_{t+h}, f_{t,h})$, and the function $c_{t+h}(\cdot)$ can change over t and the forecast horizon h.

Markov-switching model features parameters changing in different regimes, but in contrast with the threshold models the change is dictated by a non-observable state variable that is modelled as a hidden Markov chain.

Martingale property tomorrow's asset price is expected to be equal to today's price given some information set

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t .$$

Nonparametric regression is a data driven technique where a conditional moment of a random variable is specified as an unknown function of the data and estimated by means of a kernel or any other weighting scheme on the data.

Random field a scalar random field is defined as a function $m(\omega, x) : \Omega \times A \rightarrow R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$.

Sieves the sieves or approximating spaces are approximations to an unknown function, that are dense in the original function space. Sieves can be constructed using linear spans of power series, e. g., Fourier series, splines, or many other basis functions such as artificial neural network (ANN), and various polynomials (Hermite, Laguerre, etc.).

Smooth transition models threshold model with the indicator function replaced by a smooth monotonically increasing differentiable function such as a probability distribution function.

Threshold model a nonlinear model with time-varying coefficients specified by using an indicator which takes a non-zero value when a state variable falls on a specified partition of a set of states, and zero otherwise. The number of partitions is finite.

Varying cross-sectional rank (VCR) of asset i is the proportion of assets that have a return less than or equal to the return of firm i at time t

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t})$$

Volatility Volatility in financial economics is often measured by the conditional variance (e. g., ARCH) or the conditional range. It is important for any decision making under uncertainty such as portfolio allocation, option pricing, risk management.

Definition of the Subject

Financial Forecasting

Financial forecasting is concerned with the prediction of prices of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc. Though many agents in the economy, i. e. investors, money managers, investment banks, hedge funds, etc. are interested in the forecasting of financial prices per se, the importance of financial forecasting derives primarily from the role of financial markets within the macro economy. The development of financial instruments and financial institutions contribute to the growth and stability of the overall economy. Because of this interconnection between financial markets and the real economy, financial forecasting is also intimately linked to macroeconomic forecasting, which is concerned with the prediction of macroeconomic aggregates such as growth of the gross domestic product, consumption growth, inflation rates, commodities prices, etc. Financial forecasting and macroeconomic forecasting share many of the techniques and statistical models that will be explained in detail in this article.

In financial forecasting a major object of study is the return to a financial asset, mostly calculated as the continuously compounded return, i. e., $y_t = \log p_t - \log p_{t-1}$ where p_t is the price of the asset at time t. Nowadays financial forecasters use sophisticated techniques that combine the advances in modern finance theory, pioneered by Markowitz [113], with the advances in time series econometrics, in particular the development of nonlinear models for conditional moments and conditional quantiles of asset returns.

The aim of finance theory is to provide models for expected returns taking into account the uncertainty of the future asset payoffs. In general, financial models are concerned with investors' decisions under uncertainty. For instance the portfolio allocation problem deals with the allocation of wealth among different assets that carry different levels of risk. The implementation of these theories relies on econometric techniques that aim to estimate financial models and testing them against the data. Financial econometrics is the branch of econometrics that provides model-based statistical inference for financial variables, and therefore financial forecasting will provide their corresponding model-based predictions. However there are also econometric developments that inform the construction of ad hoc time series models that are valuable on describing the stylized facts of financial data.

Since returns $\{y_t\}$ are random variables, the aim of financial forecasting is to forecast conditional moments, quantiles, and eventually the conditional distribution of these variables. Most of the time our interest will be centered on expected returns and volatility as these two moments are crucial components on portfolio allocation problems, option valuation, and risk management, but it is also possible to forecast quantiles of a random variable, and therefore to forecast the expected probability density function. Density forecasting is the most complete forecast as it embeds all the information on the financial variable of interest. Financial forecasting is also concerned with other financial variables like durations between trades and directions of price changes. In these cases, it is also possible to construct conditional duration models and conditional probit models that are the basis for forecasting durations and market timing.

Critical to the understanding of the methodological development in financial forecasting is the statistical concept of *martingale*, which historically has its roots in the games of chance also associated with the beginnings of probability theory in the XVI century. Borrowing from the concept of fair game, financial prices are said to enjoy the *martingale property* if tomorrow's price is expected to be equal to today's price given some information set; in other words tomorrow's price has an equal chance to either move up or move down, and thus the best forecast must be the current price. The martingale property is written as

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t$$

where $\mathbb{E}$ is the expectation operator and the information set $\mathcal{F}_t \equiv \{p_t, p_{t-1}, p_{t-2}, \dots\}$ is the collection of past and current prices, though it may also include other variables known at time t such as volume. From a forecasting point of view, the martingale model implies that changes in financial prices $(p_{t+1} - p_t)$ are not predictable.

The most restrictive form of the martingale property, proposed by Bachelier [6] in his theory of speculation is the model (in logarithms)

$$\log p_{t+1} = \mu_t + \log p_t + \varepsilon_{t+1} ,$$

where $\mu_t = \mu$ is a constant drift and ε_{t+1} is an identically and independently distributed (i.i.d.) error that is assumed to be normally distributed with zero mean and constant variance σ^2. This model is also known as a random walk model. Since the return is the percentage change in prices, i. e. $y_t = \log p_t - \log p_{t-1}$, an equivalent model for asset returns is

$$y_{t+1} = \mu_t + \varepsilon_{t+1} .$$

Then, taking conditional expectations, we find that $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_t$. If the conditional mean return is not time-varying, $\mu_t = \mu$, then the returns are not forecastable based on past price information. In addition and given the assumptions on the error term, returns are independent and identically distributed random variables. These two properties, a constant drift and an i.i.d error term, are too restrictive and they rule out the possibility of any predictability in asset returns. A less restrictive and more plausible version is obtained when the i.i.d assumption is relaxed. The error term may be heteroscedastic so that returns have different (unconditional or conditional) variances and consequently they are not identically distributed, and/or the error term, though uncorrelated, may exhibit dependence in higher moments and in this case the returns are not independent random variables.

The advent of modern finance theory brings the notion of systematic risk, associated with return variances and covariances, into asset pricing. Though these theories were developed to explain the cross-sectional variability of financial returns, they also helped many years later with the construction of time series models for financial returns. Arguably, the two most important asset pricing models in modern finance theory are the Capital Asset Pricing Model (CAPM) proposed by Sharpe [137] and Lintner [103] and the Arbitrage Pricing Theory (APT) proposed by Ross [131]. Both models claim that the expected return to an asset is a linear function of risk; in CAPM risk is related to the covariance of the asset return with the return to the market portfolio, and in APT risk is measured as exposure to a set of factors, which may include the market portfolio among others. The original version of CAPM, based on the assumption of normally distributed returns, is written as

$$\mathbb{E}(y_i) = y_f + \beta_{im}\left[\mathbb{E}(y_m) - y_f\right] ,$$

where y_f is the risk-free rate, y_m is the return to the market portfolio, and β_{im} is the risk of asset i defined as

$$\beta_{im} = \frac{\operatorname{cov}(y_i, y_m)}{\operatorname{var}(y_m)} = \frac{\sigma_{im}}{\sigma_m^2} .$$

This model has a time series version known as the conditional CAPM [17] that it may be useful for forecasting purposes. For asset i and given an information set as $\mathcal{F}_t = \{y_{i,t}, y_{i,t-1}, \ldots ; y_{m,t}, y_{m,t-1}, \ldots\}$, the expected return is a linear function of a time-varying beta

$$\mathbb{E}(y_{i,t+1}|\mathcal{F}_t) = y_f + \beta_{im,t}\left[\mathbb{E}(y_{m,t+1}|\mathcal{F}_t) - y_f\right]$$

where $\beta_{im,t} = \frac{\operatorname{cov}(y_{i,t+1}, y_{m,t+1}|\mathcal{F}_t)}{\operatorname{var}(y_{m,t+1}|\mathcal{F}_t)} = \frac{\sigma_{im,t}}{\sigma_{m,t}^2}$. From this type of models is evident that we need to model the conditional second moments of returns jointly with the conditional mean. A general finding of this type of models is that when there is high volatility, expected returns are high, and hence forecasting volatility becomes important for the forecasting of expected returns. In the same spirit, the APT models have also conditional versions that exploit the information contained in past returns. A K-factor APT model is written as

$$y_t = c + B' f_t + \varepsilon_t ,$$

where f_t is a $K\times 1$ vector of factors and B is a $K\times 1$ vector of sensitivities to the factors. If the factors have time-varying second moments, it is possible to specify an APT model with a factor structure in the time-varying covariance matrix of asset returns [48], which in turn can be exploited for forecasting purposes.

The conditional CAPM and conditional APT models are fine examples on how finance theory provides a base to specify time-series models for financial returns. However there are other time series specifications, more *ad hoc* in nature, that claim that financial prices are nonlinear functions – not necessarily related to time-varying second moments – of the information set and by that, they impose some departures from the martingale property. In this case it is possible to observe some predictability in asset prices. This is the subject of nonlinear financial forecasting. We begin with a precise definition of linearity versus nonlinearity.

Linearity and Nonlinearity

Lee, White, and Granger [99] are the first who precisely define the concept of "linearity". Let $\{Z_t\}$ be a stochastic process, and partition Z_t as $Z_t = (y_t \; X_t')'$, where (for simplicity) y_t is a scalar and X_t is a $k \times 1$ vector. X_t may (but need not necessarily) contain a constant and lagged values of y_t. LWG define that the process $\{y_t\}$ is *linear in mean conditional on X_t* if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k .$$

In the context of forecasting, Granger and Lee [71] define linearity as follows. Define $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$ being the optimum least squares h-step forecast of y_{t+h} made at time t. μ_{t+h} will generally be a nonlinear function of the contents of $\mathcal{F}_t$. Denote m_{t+h} the optimum *linear* forecast of y_{t+h} made at time t be the best forecast that is constrained to be a linear combination of the contents of $X_t \in \mathcal{F}_t$. Granger and Lee [71] define that $\{y_t\}$ is said to be *linear in conditional mean* if μ_{t+h} is linear in X_t, i. e., $\Pr\left[\mu_{t+h} = m_{t+h}\right] = 1$ for all t and for all h. Under this definition the focus is the conditional mean and thus a process exhibiting autoregressive conditional heteroskedasticity (ARCH) [44] may nevertheless exhibit linearity of this sort because ARCH does not refer to the conditional mean. This is appropriate whenever we are concerned with the adequacy of linear models for forecasting the conditional mean returns. See [161], Section 2, for a more rigorous treatment on the definitions of linearity and nonlinearity.

This definition may be extended with some caution to the concept of linearity in higher moments and quantiles, but the definition may depend on the focus or interest of the researcher. Let $\varepsilon_{t+h} = y_{t+h} - \mu_{t+h}$ and $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$. If we consider the ARCH and GARCH as linear models, we say $\left\{\sigma^2_{t+h}\right\}$ is linear in conditional variance if σ^2_{t+h} is a linear function of lagged ε^2_{t-j} and σ^2_{t-j} for some h or for all h. Alternatively, $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$ is said to be linear in conditional variance if σ^2_{t+h} is a linear function of $x_t \in \mathcal{F}_t$ for some h or for all h. Similarly, we may consider linearity in conditional quantiles. The issue of linearity versus nonlinearity is most relevant for the conditional mean. It is more relevant whether a certain specification is correct or incorrect (rather than linear or nonlinear) for higher order conditional moments or quantiles.

Introduction

There exists a nontrivial gap between martingale difference and serial uncorrelatedness. The former implies the latter, but not vice versa. Consider a stationary time series $\{y_t\}$. Often, serial dependence of $\{y_t\}$ is described by its autocorrelation function $\rho(j)$, or by its standardized spectral density

$$h(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \rho(j) \mathrm{e}^{-ij\omega}, \quad \omega \in [-\pi, \pi] \, .$$

Both $h(\omega)$ and $\rho(j)$ are the Fourier transform of each other, containing the same information of serial correlations of $\{y_t\}$. A problem with using $h(\omega)$ and $\rho(j)$ is that they cannot capture nonlinear time series that have zero autocorrelation but are not serially independent. Nonlinear MA and Bilinear series are good examples:

$$\text{Nonlinear MA}: \quad Y_t = b e_{t-1} e_{t-2} + e_t \, ,$$

$$\text{Bilinear}: \quad Y_t = b e_{t-1} Y_{t-2} + e_t \, .$$

These processes are serially uncorrelated, but they are predictable using the past information. Hong and Lee [86] note that the autocorrelation function, the variance ratios, and the power spectrum can easily miss these processes. Misleading conclusions in favor of the martingale hypothesis could be reached when these test statistics are insignificant. It is therefore important and interesting to explore whether there exists a gap between serial uncorrelatedness and martingale difference behavior for financial forecasting, and if so, whether the neglected nonlinearity in conditional mean can be explored to forecast financial asset returns.

In the forthcoming sections, we will present, without being exhaustive, nonlinear time series models for financial returns, which are the basis for nonlinear forecasting. In Sect. "Nonlinear Forecasting Models for the Conditional Mean", we review nonlinear models for the conditional mean of returns. A general representation is $y_{t+1} = \mu(y_t, y_{t-1}, \ldots) + \varepsilon_{t+1}$ with $\mu(\cdot)$ a nonlinear function of the information set. If $\mathbb{E}(y_{t+1}|y_t, y_{t-1}, \ldots) = \mu(y_t, y_{t-1}, \ldots)$, then there is a departure from the martingale hypothesis, and past price information will be relevant to predict tomorrow's return. In Sect. "Nonlinear Forecasting Models for the Conditional Variance", we review models for the conditional variance of returns. For instance, a model like $y_{t+1} = \mu + u_{t+1}\sigma_{t+1}$ with time-varying conditional variance $\sigma^2_{t+1} = \mathbb{E}((y_{t+1} - \mu)^2|\mathcal{F}_t)$ and i.i.d. error u_{t+1}, is still a martingale-difference for returns but it represents a departure from the independence assumption. The conditional mean return may not be predictable but the conditional variance of the return will be. In addition, as we have seen modeling time-varying variances and covariances will be very useful for the implementation of conditional CAPM and APT models.

Nonlinear Forecasting Models for the Conditional Mean

We consider models to forecast the expected price changes of financial assets and we restrict the loss function of the forecast error to be the mean squared forecast error (MSFE). Under this loss, the optimal forecast is $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$. Other loss functions may also be used but it will be necessary to forecast other aspects of the forecast density. For example, under a mean absolute error loss function the optimal forecast is the conditional median.

There is evidence for μ_{t+h} being time-varying. Simple linear autoregressive polynomials in lagged price changes are not sufficient to model μ_{t+h} and nonlinear specifications are needed. These can be classified into parametric and nonparametric. Examples of parametric models are autoregressive bilinear and threshold models. Examples of nonparametric models are artificial neural network, kernel and nearest neighbor regression models.

It will be impossible to have an exhaustive review of the many nonlinear specifications. However, as discussed in White [161] and Chen [25], some nonlinear models are universal approximators. For example, the sieves or approximating spaces are proven to approximate very well unknown functions and they can be constructed using linear spans of power series, Fourier series, splines, or many other basis functions such as artificial neural network (ANN), Hermite polynomials as used in e. g., [56] for modelling semi-nonparametric density, and Laguerre polynomials used in [119] for modelling the yield curve. Diebold and Li [36] and Huang, Lee, and Li [89] use the Nelson–Siegel model in forecasting yields and inflation.

We review parametric nonlinear models like threshold model, smooth transition model, Markov switching model, and random fields model; nonparametric models like local linear, local polynomial, local exponential, and functional coefficient models; and nonlinear models based on sieves like ANN and various polynomials approximations. For other nonlinear specifications we recommend some books on nonlinear time series models such as Fan and Yao [52], Gao [57], and Tsay [153]. We begin with a very simple nonlinear model.

A Simple Nonlinear Model with Dummy Variables

Goyal and Welch [66] forecast the equity premium on the S&P 500 index – index return minus T-bill rate – using many predictors such as stock-related variables (e. g., dividend-yield, earning-price ratio, book-to-market ratio, corporate issuing activity, etc.), interest-rate-related variables (e. g., treasury bills, long-term yield, corporate bond returns, inflation, investment to capital ratio), and ex ante consumption, wealth, income ratio (modified from [101]). They find that these predictors have better performance in bad times, such as the Great Depression (1930–33), the oil-shock period (1973–75), and the tech bubble-crash period (1999–2001). Also, they argue that it is reasonable to impose a lower bound (e. g., zero or 2%) on the equity premium because no investor is interested in (say) a negative premium.

Campbell and Thompson [23], inspired by the out-of-sample forecasting of Goyal and Welch [66], argue that if we impose some restrictions on the signs of the predictors' coefficients and excess return forecasts, some predictors can beat the historical average equity premium. Similarly to Goyal and Welch [66], they also use a rich set of forecasting variables – valuation ratios (e. g., dividend price ratio, earning price ratio, and book to market ratio), real return on equity, nominal interest rates and inflation, and equity share of new issues and consumption-wealth ratio. They impose two restrictions – the first one is to restrict the predictors' coefficients to have the theoretically expected sign and to set wrong-signed coefficients to zero, and the second one is to rule out a negative equity premium forecast. They show that the effectiveness of these theoretically-inspired restrictions almost always improve the out-of sample performance of the predictive regressions. This is an example where "shrinkage" works, that is to reduce the forecast error variance at the cost of a higher forecast bias but with an overall smaller mean squared forecast error (the sum of error variance and the forecast squared bias).

The results from Goyal and Welch [66] and Campbell and Thompson [23] support a simple form of nonlinearity that can be generalized to threshold models or time-varying coefficient models, which we consider next.

Threshold Models

Many financial and macroeconomic time series exhibit different characteristics over time depending upon the state of the economy. For instance, we observe bull and bear stock markets, high volatility versus low volatility periods, recessions versus expansions, credit crunch versus excess liquidity, etc. If these different regimes are present in economic time series data, econometric specifications should go beyond linear models as these assume that there is only a single structure or regime over time. Nonlinear time series specifications that allow for the possibility of different regimes, also known as state-dependent models, include several types of models: threshold, smooth transition, and regime-switching models.

Threshold autoregressive (TAR) models [148,149] assume that the dynamics of the process is explained by an autoregression in each of the n regimes dictated by a conditioning or threshold variable. For a process $\{y_t\}$, a general specification of a TAR model is

$$y_t = \sum_{j=1}^{n} \left[\phi_o^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} y_{t-i} + \varepsilon_t^{(j)} \right] \mathbf{1}(r_{j-1} < x_t \leq r_j) .$$

There are n regimes, in each one there is an autoregressive process of order p_j with different autoregressive parameters $\phi_i^{(j)}$, the threshold variable is x_t with r_j thresholds and $r_o = -\infty$ and $r_n = +\infty$, and the error term is assumed i.i.d. with zero mean and different variance across regimes $\varepsilon_t^{(j)} \sim \text{i.i.d.}\left(0, \sigma_j^2\right)$, or more generally $\varepsilon_t^{(j)}$ is assumed to be a martingale difference. When the threshold variable is the lagged dependent variable itself y_{t-d}, the model is known as self-exciting threshold autoregressive (SETAR) model. The SETAR model has been applied to the modelling of exchange rates, industrial production indexes, and gross national product (GNP) growth, among other economic data sets. The most popular specifications within economic time series tend to find two, at most three regimes. For instance, Boero and Marrocu [18] compare a two and three-regime SETAR models with a linear AR with GARCH disturbances for the euro exchange rates. On the overall forecasting sample, the linear model performs better than the SETAR models but there is some improvement in the predictive performance of the SETAR model when conditioning on the regime.

Smooth Transition Models

In the SETAR specification, the number of regimes is discrete and finite. It is also possible to model a *continuum* of regimes as in the Smooth Transition Autoregressive (STAR) models [144]. A typical specification is

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \left(\theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i}\right) F(y_{t-d}) + \varepsilon_t$$

where $F(y_{t-d})$ is the transition function that is continuous and in most cases is either a logistic function or an exponential,

$$F(y_{t-d}) = \left[1 + \exp\left(-\gamma\left(y_{t-d} - r\right)\right)\right]^{-1}$$
$$F(y_{t-d}) = 1 - \left[\exp\left(-\gamma\left(y_{t-d} - r\right)^2\right)\right]$$

This model can be understood as many autoregressive regimes dictated by the values of the function $F(y_{t-d})$, or alternatively as an autoregression where the autoregressive parameters change smoothly over time. When $F(y_{t-d})$ is logistic and $\gamma \to \infty$, the STAR model collapses to a threshold model SETAR with two regimes. One important characteristic of these models, SETAR and STAR, is that the process can be stationary within some regimes and non-stationary within others moving between explosive and contractionary stages.

Since the estimation of these models can be demanding, the first question to solve is whether the nonlinearity is granted by the data. A test for linearity is imperative before engaging in the estimation of nonlinear specifications. An LM test that has power against the two alternatives specifications SETAR and STAR is proposed by Luukkonen et al. [110] and it consists of running two regressions: under the null hypothesis of linearity, a linear autoregression of order p is estimated in order to calculate the sum of squared residuals, SSE_0; the second is an auxiliary regression

$$\begin{aligned} y_t = \beta_0 &+ \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p}\sum_{j=1}^{p} \psi_{ij} y_{t-i} y_{t-j} \\ &+ \sum_{i=1}^{p}\sum_{j=1}^{p} \zeta_{ij} y_{t-i} y_{t-j}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} \xi_{ij} y_{t-i} y_{t-j}^3 + u_t \end{aligned}$$

from which we calculate the sum of squared residuals, SSE_1. The test is constructed as $\chi^2 = T(SSE_0 - SSE_1)/SSE_0$ that under the null hypothesis of linearity is chi-squared distributed with $p(p+1)/2 + 2p^2$ degrees of freedom. There are other tests in the literature, for instance Hansen [80] proposes a likelihood ratio test that has a non-standard distribution, which is approximated by implementing a bootstrap procedure. Tsay [151] proposes a test based on arranged regressions with respect to the increasing order of the threshold variable and by doing this the testing problem is transformed into a change-point problem.

If linearity is rejected, we proceed with the estimation of the nonlinear specification. In the case of the SETAR model, if we fix the values of the delay parameter d and the thresholds r_j, the model reduces to n linear regressions for which least squares estimation is straightforward. Tsay [151] proposes a conditional least squares (CLS) estimator. For simplicity of exposition suppose that there are two regimes in the data and the model to estimate is

$$\begin{aligned} y_t = &\left[\phi_o^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} y_{t-i}\right] \mathbf{1}(y_{t-d} \le r) \\ &+ \left[\phi_o^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} y_{t-i}\right] \mathbf{1}(y_{t-d} > r) + \varepsilon_t \end{aligned}$$

Since r and d are fixed, we can apply least squares estimation to the model and to obtain the LS estimates for the parameters ϕ_i's. With the LS residual $\hat{\varepsilon}_t$, we obtain the total sum of squares $S(r,d) = \sum_t \hat{\varepsilon}_t^2$. The CLS estimates of r and d are obtained from $(\hat{r}, \hat{d}) = \arg\min S(r,d)$.

For the STAR model, it is also necessary to specify a priori the functional form of $F(y_{t-d})$. Teräsvirta [144] proposes a modeling cycle consisting of three stages: specification, estimation, and evaluation. In general, the specification stage consists of sequence of null hypothesis to be tested within a linearized version of the STAR model.

Parameter estimation is carried out by nonlinear least squares or maximum likelihood. The evaluation stage mainly consists of testing for no error autocorrelation, no remaining nonlinearity, and parameter constancy, among other tests.

Teräsvirta and Anderson [146] find strong nonlinearity in the industrial production indexes of most of the OECD countries. The preferred model is the logistic STAR with two regimes, recessions and expansions. The dynamics in each regime are country dependent. For instance, in USA they find that the economy tends to move from recessions into expansions very aggressively but it will take a large negative shock to move rapidly from an expansion into a recession. Other references for applications of these models to financial series are found in [28,73,94].

For forecasting with STAR models, see Lundbergh and Teräsvirta [109]. It is easy to construct the one-step-ahead forecast but the multi-step-ahead forecast is a complex problem. For instance, for the 2-regime threshold model, the one-step-ahead forecast is constructed as the conditional mean of the process given some information set

$$\begin{aligned}&\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta)\\ &= \left[\phi_o^{(1)} + \sum_{i=1}^{p_1}\phi_i^{(1)} y_{t+1-i}\right]\mathbf{1}(y_{t+1-d} \le r)\\ &\quad + \left[\phi_o^{(2)} + \sum_{i=1}^{p_2}\phi_i^{(2)} y_{t+1-i}\right]\mathbf{1}(y_{t+1-d} > r)\end{aligned}$$

provided that $y_{t+1-i}, y_{t+1-d} \in \mathcal{F}_t$. However, a multi-step-ahead forecast will be a function of variables that being dated at a future date do not belong to the information set; in this case the solution requires the use of numerical integration techniques or simulation/bootstrap procedures. See Granger and Teräsvirta [72], Chapter 9, and Teräsvirta [145] for more details on numerical methods for multi-step forecasts.

Markov-Switching Models

A Markov-switching (MS) model [76,77] also features changes in regime, but in contrast with the SETAR models the change is dictated by a non-observable state variable that is modelled as a Markov chain. For instance, a first order autoregressive Markov switching model is specified as

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \varepsilon_t$$

where $s_t = 1, 2, \dots, N$ is the unobserved state variable that is modelled as an N-state Markov chain with transition probabilities $p_{ij} = P(s_t = j|s_{t-1} = i)$, and $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$ or more generally ε_t is a martingale difference. Conditioning in a given state and an information set $\mathcal{F}_t$, the process $\{y_t\}$ is linear but unconditionally the process is nonlinear. The conditional forecast is $\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t; \theta) = c_j + \phi_j y_t$ and the unconditional forecast based on observable variables is the sum of the conditional forecasts for each state weighted by the probability of being in that state,

$$\begin{aligned}&\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta)\\ &= \sum_{j=1}^{N} P(s_{t+1} = j|\mathcal{F}_t;\theta)\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t;\theta)\,.\end{aligned}$$

The parameter vector $\theta = (c_1 \dots c_N, \phi_1 \dots \phi_N, \sigma^2)'$ as well as the transition probabilities p_{ij} can be estimated by maximum likelihood.

MS models have been applying to the modeling of foreign exchange rates with mixed success. Engel and Hamilton [43] fit a two-state MS for the Dollar and find that there are long swings and by that they reject the random walk behavior in the exchange rate. Marsh [114] estimates a two-state MS for the Deutschemark, the Pound Sterling, and the Japanese Yen. Though the model approximates the characteristics of the data well, the forecasting performance is poor when measured by the profit/losses generated by a set of trading rules based on the predictions of the MS model. On the contrary, Dueker and Neely [40] find that for the same exchange rate a MS model with three states variables – in the scale factor of the variance of a Student-t error, in the kurtosis of the error, and in the expected return – produces out-of-sample excess returns that are slightly superior to those generated by common trading rules. For stock returns, there is evidence that MS models perform relatively well on describing two states in the mean (high/low returns) and two states in the variance (stable/volatile periods) of returns [111]. In addition, Perez-Quiros and Timmermann [124] propose that the error term should be modelled as a mixture of Gaussian and Student-t distributions to capture the outliers commonly found in stock returns. This model provides some gains in predictive accuracy mainly for small firms returns. For interest rates in USA, Germany, and United Kingdom, Ang and Bekaert [5] find that a two-state MS model that incorporates information on international short rate and on term spread is able to predict better than an univariate MS model. Additionally they find that in USA the classification of regimes correlates well with the business cycles.

SETAR, STAR, and MS models are successful specifications to approximate the characteristics of financial and macroeconomic data. However, good in-sample performance does not imply necessarily a good out-of-sample performance, mainly when compared to simple linear ARMA models. The success of nonlinear models depends on how prominent the nonlinearity is in the data. We should not expect a nonlinear model to perform better than a linear model when the contribution of the nonlinearity to the overall specification of the model is very small. As it is argued in Granger and Teräsvirta [72], the prediction errors generated by a nonlinear model will be smaller only when the nonlinear feature modelled in-sample is also present in the forecasting sample.

A State Dependent Mixture Model Based on Cross-sectional Ranks

In the previous section, we have dealt with nonlinear time series models that only incorporate time series information. González-Rivera, Lee, and Mishra [63] propose a nonlinear model that combines time series with cross sectional information. They propose the modelling of expected returns based on the joint dynamics of a sharp jump in the cross-sectional rank and the realized returns. They analyze the marginal probability distribution of a jump in the cross-sectional rank within the context of a duration model, and the probability of the asset return conditional on a jump specifying different dynamics depending on whether or not a jump has taken place. The resulting model for expected returns is a mixture of normal distributions weighted by the probability of jumping.

Let $y_{i,t}$ be the return of firm i at time t, and $\{y_{i,t}\}_{i=1}^{M}$ be the collection of asset returns of the M firms that constitute the *market* at time t. For each time t, the asset returns are ordered from the smallest to the largest, and define $z_{i,t}$, the *Varying Cross-sectional Rank* (VCR) of firm i within the market, as the proportion of firms that have a return less than or equal to the return of firm i. We write

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t}) \,, \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and for M large, $z_{i,t} \in (0, 1]$. Since the rank is a highly dependent variable, it is assumed that small movements in the asset ranking will not contain significant information and that most likely large movements in ranking will be the result of news in the overall market and/or of news concerning a particular asset. Focusing on large rank movements, we define, at time t, a sharp jump as a binary variable that takes the value one when there is a minimum (upward or downward) movement of 0.5 in the ranking of asset i, and zero otherwise:

$$J_{i,t} \equiv \mathbf{1}(|z_{i,t} - z_{i,t-1}| \geq 0.5) \,. \tag{2}$$

A jump of this magnitude brings the asset return above or below the median of the cross-sectional distribution of returns. Note that this notion of jumps differs from the more traditional meaning of the word in the context of continuous-time modelling of the univariate return process. A jump in the cross-sectional rank implicitly depends on numerous univariate return processes.

The analytical problem now consists in modeling the joint distribution of the return $y_{i,t}$ and the jump $J_{i,t}$, i.e. $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1})$ where $\mathcal{F}_{t-1}$ is the information set up to time $t-1$. Since $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1}) = f_1(J_{i,t}|\mathcal{F}_{t-1}) f_2(y_{i,t}|J_{i,t}, \mathcal{F}_{t-1})$, the analysis focuses first on the modelling of the marginal distribution of the jump, and subsequently on the modelling of the conditional distribution of the return.

Since $J_{i,t}$ is a Bernoulli variable, the marginal distribution of the jump is $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1 - p_{i,t})^{(1-J_{i,t})}$ where $p_{i,t} \equiv \Pr(J_{i,t} = 1|\mathcal{F}_{t-1})$ is the conditional probability of a jump in the cross-sectional ranks. The modelling of $p_{i,t}$ is performed within the context of a dynamic duration model specified in calendar time as in Hamilton and Jordà [79]. The calendar time approach is necessary because asset returns are reported in calendar time (days, weeks, etc.) and it has the advantage of incorporating any other available information also reported in calendar time.

It is easy to see that the probability of jumping and duration must have an inverse relationship. If the probability of jumping is high, the expected duration must be short, and vice versa. Let $\Psi_{N(t)}$ be the expected duration. The expected duration until the next jump in the cross-sectional rank is given by $\Psi_{N(t)} = \sum_{j=1}^{\infty} j(1 - p_t)^{j-1} p_t = p_t^{-1}$. Note that $\sum_{j=0}^{\infty}(1 - p_t)^j = p_t^{-1}$. Differentiating with respect to p_t yields $\sum_{j=0}^{\infty} -j(1 - p_t)^{j-1} = -p_t^{-2}$. Multiplying by $-p_t$ gives $\sum_{j=0}^{\infty} j(1 - p_t)^{j-1} p_t = p_t^{-1}$ and thus $\sum_{j=1}^{\infty} j(1 - p_t)^{j-1} p_t = p_t^{-1}$. Consequently, to model $p_{i,t}$, it suffices to model the expected duration and compute its inverse. Following Hamilton and Jordà [79], an autoregressive conditional hazard (ACH) model is specified. The ACH model is a calendar-time version of the autoregressive conditional duration (ACD) of Engle and Russell [49]. In both ACD and ACH models, the expected duration is a linear function of lag durations. However as the ACD model is set up in event time, there are some difficulties on how to introduce information that arrives between events. This is not the case in the ACH model because the set-up is in calendar time. In the ACD model, the forecasting object is the expected time between events; in the ACH model, the objective is to forecast the probability that the event will happen tomorrow given the information known up to today. A general ACH model is specified as

$$\Psi_{N(t)} = \sum_{j=1}^{m} \alpha_j D_{N(t)-j} + \sum_{j=1}^{r} \beta_j \Psi_{N(t)-j} \,. \tag{3}$$

Since p_t is a probability, it must be bounded between zero and one. This implies that the conditional duration must have a lower bound of one. Furthermore, working in calendar time it is possible to incorporate information that becomes available between jumps and can affect the probability of a jump in future periods. The conditional hazard rate is specified as

$$p_t = [\Psi_{N(t-1)} + \delta' X_{t-1}]^{-1} \,, \tag{4}$$

where X_{t-1} is a vector of relevant calendar time variables such as past VCRs and past returns. This completes the marginal distribution of the jump $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1-p_{i,t})^{(1-J_{i,t})}$.

On modelling $f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$, it is assumed that the return to asset i may behave differently depending upon the occurrence of a jump. The modelling of two potential different states (whether a jump has occurred or not) will permit to differentiate whether the conditional expected return is driven by active or/and passive movements in the asset ranking in conjunction with its own return dynamics. A priori, different dynamics are possible in these two states. A general specification is

$$f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) = \begin{cases} N(\mu_{1,t}, \sigma_{1,t}^2) & \text{if } J_t = 1 \\ N(\mu_{0,t}, \sigma_{0,t}^2) & \text{if } J_t = 0 \end{cases} \,, \tag{5}$$

where $\mu_{j,t}$ is the conditional mean and $\sigma_{j,t}^2$ the conditional variance in each state ($j = 1, 0$). Whether these two states are present in the data is an empirical question and it should be answered through statistical testing.

Combining the models for the marginal density of the jump and the conditional density of the returns, the estimation can be conducted with maximum likelihood techniques. For a sample $\{y_t, J_t\}_{t=1}^{T}$, the joint log-likelihood function is

$$\begin{aligned}&\sum_{t=1}^{T} \ln f(y_t, J_t|\mathcal{F}_{t-1}; \theta)\\ &= \sum_{t=1}^{T} \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1) + \sum_{t=1}^{T} \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) \,.\end{aligned}$$

Let us call $\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1)$ and $\mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$. The maximization of the joint log-likelihood function can be achieved by maximizing $\mathcal{L}_1(\theta_1)$ and $\mathcal{L}_2(\theta_2)$ separately without loss of efficiency by assuming that the parameter vectors θ_1 and θ_2 are "variation free" in the sense of Engle et al. [45].

The log-likelihood function $\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \ln f_1(J_t| \mathcal{F}_{t-1}; \theta_1)$ is

$$\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \left[J_t \ln p_t(\theta_1) + (1 - J_t) \ln(1 - p_t(\theta_1)) \right] , \tag{6}$$

where θ_1 includes all parameters in the conditional duration model.

The log-likelihood function $\mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln f_2(y_t| J_t, \mathcal{F}_{t-1}; \theta_2)$ is

$$\begin{aligned}\mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln \Bigg[&\frac{J_t}{\sqrt{2\pi\sigma_{1,t}^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_t - \mu_{1,t}}{\sigma_{1,t}} \right)^2 \right\} \\ &+ \frac{1 - J_t}{\sqrt{2\pi\sigma_{0,t}^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_t - \mu_{0,t}}{\sigma_{0,t}} \right)^2 \right\} \Bigg] ,\end{aligned}$$

where θ_2 includes all parameters in the conditional means and conditional variances under both regimes.

If the two proposed states are granted in the data, the marginal density function of the asset return must be a mixture of two normal density functions where the mixture weights are given by the probability of jumping p_t:

$$\begin{aligned} g(y_t|\mathcal{F}_{t-1}; \theta) &\equiv \sum_{J_t=0}^{1} f(y_t, J_t|\mathcal{F}_{t-1}; \theta) \\ &= \sum_{J_t=0}^{1} f_1(J_t|\mathcal{F}_{t-1}; \theta_1) f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) \\ &= p_t \cdot f_2(y_t|J_t = 1, \mathcal{F}_{t-1}; \theta_2) \\ &\quad + (1 - p_t) \cdot f_2(y_t|J_t = 0, \mathcal{F}_{t-1}; \theta_2), \end{aligned} \tag{7}$$

as $f_1(J_t|\mathcal{F}_{t-1}; \theta_1) = p_t^{J_t}(1 - p_t)^{(1-J_t)}$. Therefore, the one-step ahead forecast of the return is

$$\begin{aligned} &\mathbb{E}(y_{t+1}|\mathcal{F}_t; \theta) \\ &= \int y_{t+1} \cdot g(y_{t+1}|\mathcal{F}_t; \theta) \mathrm{d}y_{t+1} \\ &= p_{t+1}(\theta_1) \cdot \mu_{1,t+1}(\theta_2) + (1 - p_{t+1}(\theta_1)) \cdot \mu_{0,t+1}(\theta_2) \,. \end{aligned} \tag{8}$$

The expected return is a function of the probability of jumping p_t, which is a nonlinear function of the information set as shown in (4). Hence the expected returns are nonlinear functions of the information set, even in a simple case where $\mu_{1,t}$ and $\mu_{0,t}$ are linear.

This model was estimated for the returns of the constituents of the SP500 index from 1990 to 2000, and its performance was assessed in an out-of-sample exercise from

2001 to 2005 within the context of several trading strategies. Based on the one-step-ahead forecast of the mixture model, a proposed trading strategy called VCR-Mixture Trading Rule is shown to be a superior rule because of its ability to generate large risk-adjusted mean returns when compared to other technical and model-based trading rules. The VCR-Mixture Trading Rule is implemented by computing for each firm in the SP500 index the one-step ahead forecast of the return as in (8). Based on the forecasted returns $\{\hat{y}_{i,t+1}(\hat{\theta}_t)\}_{t=R}^{T-1}$, the investor predicts the VCR of all assets in relation to the overall market, that is,

$$\hat{z}_{i,t+1} = M^{-1}\sum_{j=1}^{M}\mathbf{1}(\hat{y}_{j,t+1} \leq \hat{y}_{i,t+1}), \qquad t = R, \dots, T-1\,, \tag{9}$$

and buys the top K performing assets if their forecasted return is above the risk-free rate. In every subsequent out-of-sample period ($t = R, \dots, T-1$), the investor revises her portfolio, selling the assets that fall out of the top performers and buying the ones that rise to the top, and she computes the one-period portfolio return

$$\pi_{t+1} = K^{-1}\sum_{j=1}^{M} y_{j,t+1}\cdot\mathbf{1}\left(\hat{z}_{j,t+1} \geq z_{t+1}^{K}\right), \qquad t = R, \dots, T-1\,, \tag{10}$$

where z_{t+1}^{K} is the cutoff cross-sectional rank to select the K best performing stocks such that $\sum_{j=1}^{M}\mathbf{1}\left(\hat{z}_{j,t+1} \geq z_{t+1}^{K}\right) = K$. In the analysis of González-Rivera, Lee, and Mishra [63] a portfolio is formed with the top 1% ($K = 5$ stocks) performers in the SP500 index. Every asset in the portfolio is weighted equally. The evaluation criterion is to compute the "mean trading return" over the forecasting period

$$MTR = P^{-1}\sum_{t=R}^{T-1}\pi_{t+1}\,.$$

It is also possible to correct MTR according to the level of risk of the chosen portfolio. For instance, the traditional Sharpe ratio will provide the excess return per unit of risk measured by the standard deviation of the selected portfolio

$$SR = P^{-1}\sum_{t=R}^{T-1}\frac{(\pi_{t+1} - r_{f,t+1})}{\sigma_{t+1}^{\pi}(\hat{\theta}_t)}\,,$$

where $r_{f,t+1}$ is the risk free rate. The VCR-Mixture Trading Rule produces a weekly MTR of 0.243% (63.295% cumulative return over 260 weeks), equivalent to a yearly compounded return of 13.45%, that is significantly more than the next most favorable rule, which is the Buy-and-Hold-the-Market Trading Rule with a weekly mean return of −0.019%, equivalent to a yearly return of −1.00%. To assess the return-risk trade off, we implement the Sharpe ratio. The largest SR (mean return per unit of standard deviation) is provided by the VCR-Mixture rule with a weekly return of 0.151% (8.11% yearly compounded return per unit of standard deviation), which is lower than the mean return provided by the same rule under the MTR criterion, but still a dominant return when compared to the mean returns provided by the Buy-and-Hold-the-Market Trading Rule.

Random Fields

Hamilton [78] proposed a flexible parametric regression model where the conditional mean has a linear parametric component and a potential nonlinear component represented by an isotropic Gaussian random field. The model has a nonparametric flavor because no functional form is assumed but, nevertheless, the estimation is fully parametric.

A scalar random field is defined as a function $m(\omega, x): \Omega \times A \rightarrow R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$. A random field is also denoted as $m(x)$. If $m(x)$ is a system of random variables with finite dimensional Gaussian distributions, then the scalar random field is said to be Gaussian and it is completely determined by its mean function $\mu(x) = \mathbb{E}\left[m(x)\right]$ and its covariance function with typical element $C(x, z) = \mathbb{E}\left[(m(x) - \mu(x))(m(z) - \mu(z))\right]$ for any $x, z \in A$. The random field is said to be homogeneous or stationary if $\mu(x) = \mu$ and the covariance function depends only on the difference vector $x - z$ and we should write $C(x, z) = C(x - z)$. Furthermore, the random field is said to be isotropic if the covariance function depends on $d(x, z)$, where $d(\cdot)$ is a scalar measure of distance. In this situation we write $C(x, z) = C(d(x, z))$.

The specification suggested by Hamilton [78] can be represented as

$$y_t = \beta_0 + x_t'\beta_1 + \lambda m(g \odot x_t) + \epsilon_t\,, \tag{11}$$

for $y_t \in R$ and $x_t \in R^k$, both stationary and ergodic processes. The conditional mean has a linear component given by $\beta_0 + x_t'\beta_1$ and a nonlinear component given by $\lambda m(g \odot x_t)$, where $m(z)$, for any choice of z, represents a realization of a Gaussian and homogenous random field with a moving average representation; x_t could be predetermined or exogenous and is independent of $m(\cdot)$, and ϵ_t is a sequence of independent and identically distributed $N(0,\sigma^2)$ variates independent of both $m(\cdot)$ and x_t as well as of lagged values of x_t. The scalar parameter λ represents the contribution of the nonlinear part to the conditional mean, the vector $g \in R^k_{0,+}$ drives the curvature of the conditional mean, and the symbol $\odot$ denotes element-by-element multiplication.

Let H_k be the covariance (correlation) function of the random field $m(\cdot)$ with typical element defined as $H_k(x,z) = \mathbb{E}\left[m(x)m(z)\right]$. Hamilton [78] proved that the covariance function depends solely upon the Euclidean distance between x and z, rendering the random field isotropic. For any x and $z \in R^k$, the correlation between $m(x)$ and $m(z)$ is given by the ratio of the volume of the overlap of k-dimensional unit spheroids centered at x and z to the volume of a single k-dimensional unit spheroid. If the Euclidean distance between x and z is greater than two, the correlation between $m(x)$ and $m(z)$ will be equal to zero. The general expression of the correlation function is

$$H_k(h) = \begin{cases} G_{k-1}(h,1)/G_{k-1}(0,1) & \text{if } h \leq 1 \\ 0 & \text{if } h > 1 \end{cases}, \qquad (12)$$

$$G_k(h,r) = \int_h^r (r^2 - w^2)^{k/2} \mathrm{d}w\,,$$

where $h \equiv \frac{1}{2} d_{L_2}(x,z)$, and $d_{L_2}(x,z) \equiv \left[(x-z)'(x-z)\right]^{1/2}$ is the Euclidean distance between x and z.

Within the specification (11), Dahl and González-Rivera [33] provided alternative representations of the random field that permit the construction of Lagrange multiplier tests for neglected nonlinearity, which circumvent the problem of unidentified nuisance parameters under the null of linearity and, at the same time, they are robust to the specification of the covariance function associated with the random field. They modified the Hamilton framework in two directions. First, the random field is specified in the L_1 norm instead of the L_2 norm, and secondly they considered random fields that may not have a simple moving average representation. The advantage of the L_1 norm, which is exploited in the testing problem, is that this distance measure is a linear function of the nuisance parameters, in contrast to the L_2 norm which is a nonlinear function. Logically, Dahl and González-Rivera proceeded in an opposite fashion to Hamilton. Whereas Hamilton first proposed a moving average representation of the random field, and secondly, he derived its corresponding covariance function, Dahl and González-Rivera first proposed a covariance function, and secondly they inquire whether there is a random field associated with it. The proposed covariance function is

$$C_k(h^*) = \begin{cases} (1-h^*)^{2k} & \text{if } h^* \leq 1 \\ 0 & \text{if } h^* > 1 \end{cases}, \qquad (13)$$

where $h^* \equiv \frac{1}{2} d_{L_1}(x,z) = \frac{1}{2}|x-z|'1$. The function (13) is a permissible covariance, that is, it satisfies the positive semidefiniteness condition, which is $q'C_k q \geq 0$ for all $q \neq 0_T$. Furthermore, there is a random field associated with it according to the Khinchin's theorem (1934) and Bochner's theorem (1959). The basic argument is that the class of functions which are covariance functions of homogenous random fields coincides with the class of positive semidefinite functions. Hence, (13) being a positive semidefinite function must be the covariance function of a homogenous random field.

The estimation of these models is carried out by maximum likelihood. From model (11), we can write $y \sim N(X\beta, \lambda^2 C_k + \sigma^2 I_T)$ where $y = (y_1, y_2, \ldots, y_T)'$, $X_1 = (x'_1, x'_2, \ldots, x'_T)'$, $X = (\mathbf{1} : X_1)$, $\beta = (\beta_0, \beta'_1)'$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$ and σ^2 is the variance of ϵ_t. C_k is a generic covariance function associated with the random field, which could be equal to the Hamilton spherical covariance function in (12), or to the covariance in (13). The log-likelihood function corresponding to this model is

$$\ell(\beta, \lambda^2, g, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\lambda^2 C_k + \sigma^2 I_T| - \frac{1}{2}(y - X\beta)'(\lambda^2 C_k + \sigma^2 I_T)^{-1}(y - X\beta). \qquad (14)$$

The flexible regression model has been applied successfully to detect nonlinearity in the quarterly growth rate of the US real GNP [34] and in the Industrial Production Index of sixteen OECD countries [33]. This technology is able to mimic the characteristics of the actual US business cycle. The cycle is dissected according to measures of duration, amplitude, cumulation and excess cumulation of the contraction and expansion phases. In contrast to Harding and Pagan [82] who find that nonlinear models are not uniformly superior to linear ones, the flexible regression model represents a clear improvement over linear models, and it seems to capture just the right shape of the expansion phase as opposed to Hamilton [76] and Durland and McCurdy [41] models, which tend to overestimate the cumulation measure in the expansion phase. It is found that the expansion phase must have at least two subphases: an aggressive early expansion after the trough, and a moderate/slow late expansion before the peak implying the existence of an inflexion point that we date approximately around one-third into the duration of the expansion phase. This shape lends support to parametric models of the growth rate that allow for three regimes [136], as opposed to models with just two regimes (contractions and expansions). For the Industrial Production Index, testing for nonlinearity within the flexible regression framework brings similar conclusions to those in Teräsvirta and Anderson [146], who propose parametric STAR models for industrial production data. However, the tests proposed in Dahl and González-Rivera [33], which have superior performance to detect smooth transition dynamics, seem to indicate that linearity cannot be rejected in the industrial production indexes of Japan, Austria, Belgium and Sweden as opposed to the findings of Teräsvirta and Anderson.

Nonlinear Factor Models

For the last ten years forecasting using a data-rich environment has been one of the most researched topic in economics and finance, see [140,141]. In this literature, factor models are used to reduce the dimension of the data but mostly they are linear models. Bai and Ng (BN) [7] introduce a nonlinear factor model with a quadratic principal component model as a special case. First consider a simple factor model

$$x_{it} = \lambda_i' F_t + e_{it} . \tag{15}$$

By the method of principal component, the elements of $\mathbf{f}_t$ are linear combinations of elements of $\mathbf{x}_t$. The factors are estimated by minimizing the sum of squared residuals of the linear model, $x_{it} = \lambda_i F_t + e_{it}$.

The factor model in (15) assumes a linear link function between the predictor $\mathbf{x}_t$ and the latent factors F_t. BN consider a more flexible approach by a nonlinear link function $g(\cdot)$ such that

$$g(x_{it}) = \phi_i' J_t + v_{it} ,$$

where J_t are the common factors, and ϕ_i is the vector of factor loadings. BN consider $g(x_{it})$ to be x_{it} augmented by some or all of the unique cross-products of the elements of $\{x_{it}\}_{i=1}^N$. The second-order factor model is then $x_{it}^* = \phi_i' J_t + v_{it}$ where x_{it}^* is an $N^* \times 1$ vector. Estimation of J_t then proceeds by the usual method of principal components. BN consider $x_{it}^* = \{x_{it}\, x_{it}^2\}_{i=1}^N$ with $N^* = 2N$, which they call the SPC (squared principal components).

Once the factors are estimated, the forecasting equation for y_{t+h} would be

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + \varepsilon_t .$$

The forecasting equation remains linear whatever the link function g is. An alternative way of capturing nonlinearity is to augment the forecasting equation to include functions of the factors

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + a(\hat{F}_t) + \varepsilon_t ,$$

where $a(\cdot)$ is nonlinear. A simple case when $a(\cdot)$ is quadratic is referred to as PC2 (squared factors) in BN.

BN note that the PC2 is conceptually distinct from SPC. While the PC2 forecasting model allows the volatility of factors estimated by linear principal components to have predictive power for y, the SPC model allows the factors to be possibly nonlinear functions of the predictors while maintaining a linear relation between the factors and y. Ludvigson and Ng [108] found that the square of the first factor estimated from a set of financial factors (i. e., volatility of the first factor) is significant in the regression model for the mean excess returns. In contrast, factors estimated from the second moment of data (i. e., volatility factors) are much weaker predictors of excess returns.

Artificial Neural Network Models

Consider an augmented single hidden layer feedforward neural network model $f(x_t, \theta)$ in which the network output y_t is determined given input x_t as

$$\begin{aligned} y_t &= f(x_t, \theta) + \varepsilon_t \\ &= x_t\beta + \sum_{j=1}^{q} \delta_j \psi(x_t \gamma_j) + \varepsilon_t \end{aligned}$$

where $\theta = (\beta'\gamma'\delta')'$, β is a conformable column vector of connection strength from the input layer to the output layer; γ_j is a conformable column vector of connection strength from the input layer to the hidden units, $j = 1, \ldots, q$; δ_j is a (scalar) connection strength from the hidden unit j to the output unit, $j = 1, \ldots, q$; and ψ is a squashing function (e. g., the logistic squasher) or a radial basis function. Input units x send signals to intermediate hidden units, then each of hidden unit produces an activation ψ that then sends signals toward the output unit. The integer q denotes the number of hidden units added to the affine (linear) network. When $q = 0$, we have a two layer *affine* network $y_t = x_t\beta + \varepsilon_t$. Hornick, Stinchcombe and White [88] show that neural network is a nonlinear flexible functional form being capable of approximating any Borel measurable function to any desired level of accuracy provided sufficiently many hidden units are available. Stinchcombe and White [138] show that this result holds for any $\psi(\cdot)$ belonging to the class of "generically comprehensively revealing" functions. These functions are "comprehensively revealing" in the sense that they can reveal arbitrary model misspecifications $\mathbb{E}(y_t|x_t) \neq f(x_t, \theta^*)$ with non-zero probability and they are "generic" in the sense that almost any choice for γ will reveal the misspecification.

We build an artificial neural network (ANN) model based on a test for neglected nonlinearity likely to have power against a range of alternatives. See White [158] and Lee, White, and Granger [99] on the neural network test and its comparison with other specification tests. The neural network test is based on a test function $h(x_t)$ chosen as the activations of 'phantom' hidden units $\psi(x_t\Gamma_j)$, $j = 1, \dots, q$, where Γ_j are random column vectors independent of x_t. That is,

$$\mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*|\Gamma_j] = \mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*] = 0 \quad j = 1, \dots, q , \tag{16}$$

under H_0, so that

$$\mathbb{E}(\Psi_t\varepsilon_t^*) = 0 , \tag{17}$$

where $\Psi_t = (\psi(x_t\Gamma_1), \dots, \psi(x_t\Gamma_q))'$ is a phantom hidden unit activation vector. Evidence of correlation of ε_t^* with Ψ_t is evidence against the null hypothesis that y_t is linear in mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations $\psi(x_t\Gamma_j)$ would permit an improvement in network performance. Thus the tests are based on sample correlation of affine network errors with phantom hidden unit activations,

$$n^{-1}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t = n^{-1}\sum_{t=1}^{n}\Psi_t(y_t - x_t\hat{\beta}) . \tag{18}$$

Under suitable regularity conditions it follows from the central limit theorem that $n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t \xrightarrow{d} N(0, W^*)$ as $n \to \infty$, and if one has a consistent estimator for its asymptotic covariance matrix, say $\hat{W}_n$, then an asymptotic chi-square statistic can be formed as

$$\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right)'\hat{W}_n^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right) \xrightarrow{d} \chi^2(q) . \tag{19}$$

Elements of Ψ_t tend to be collinear with X_t and with themselves. Thus LWG conduct a test on $q^* < q$ principal components of Ψ_t not collinear with x_t, denoted Ψ_t^*. This test is to determine whether or not there exists some advantage to be gained by adding hidden units to the affine network. We can estimate $\hat{W}_n$ robust to the conditional heteroskedasticity, or we may use with the empirical null distribution of the statistic computed by a bootstrap procedure that is robust to the conditional heteroskedasticity, e. g., wild bootstrap.

Estimation of an ANN model may be tedious and sometimes results in unreliable estimates. Recently, White [161] proposes a simple algorithm called QuickNet, a form of "relaxed greedy algorithm" because QuickNet searches for a single best additional hidden unit based on a sequence of OLS regressions, that may be analogous to the least angular regressions (LARS) of Efron, Hastie, Johnstone, and Tibshirani [42]. The simplicity of the QuickNet algorithm achieves the benefits of using a forecasting model that is nonlinear in the predictors while mitigating the other computational challenges to the use of nonlinear forecasting methods. See White [161], Section 5, for more details on QuickNet, and for other issues of controlling for overfit and the selection of the random parameter vectors Γ_j independent of x_t.

Campbell, Lo, and MacKinlay [22], Section 12.4, provide a review of these models. White [161] reviews the research frontier in ANN models. Trippi and Turban [150] review the applications of ANNs to finance and investment.

Functional Coefficient Models

A functional coefficient model is introduced by Cai, Fan, and Yao [24] (CFY), with time-varying and state-dependent coefficients. It can be viewed as a special case of Priestley's [127] state-dependent model, but it includes the models of Tong [149], Chen and Tsay [26] and regime-switching models as special cases. Let $\{(y_t, s_t)'\}_{t=1}^n$ be a stationary process, where y_t and s_t are scalar variables. Also let $X_t \equiv (1, y_{t-1}, \dots, y_{t-d})'$. We assume

$$\mathbb{E}(y_t|\mathcal{F}_{t-1}) = a_0(s_t) + \sum_{j=1}^{d} a_j(s_t)y_{t-j} ,$$

where the $\{a_j(s_t)\}$ are the autoregressive coefficients depending on s_t, which may be chosen as a function of X_t or something else. Intuitively, the functional coefficient model is an AR process with time-varying autoregressive coefficients. The coefficient functions $\{a_j(s_t)\}$ can be estimated by local linear regression. At each point s, we approximate $a_j(s_t)$ locally by a linear function $a_j(s_t) \approx a_j + b_j(s_t - s)$, $j = 0, 1, \dots, d$, for s_t near s, where a_j and b_j are constants. The local linear estimator at point s is then given by $\hat{a}_j(s) = \hat{a}_j$, where $\{(\hat{a}_j, \hat{b}_j)\}_{j=0}^d$ minimizes the sum of local weighted squares $\sum_{t=1}^n [y_t - \mathbb{E}(y_t|\mathcal{F}_{t-1})]^2 K_h(s_t - s)$, with $K_h(\cdot) \equiv K(\cdot/h)/h$ for a given kernel function $K(\cdot)$ and bandwidth $h \equiv h_n \to 0$ as $n \to \infty$. CFY [24], p. 944, suggest to select h using a modified multi-fold "leave-one-out-type" cross-validation based on MSFE.

It is important to choose an appropriate smooth variable s_t. Knowledge on data or economic theory may be helpful. When no prior information is available, s_t may be chosen as a function of explanatory vector X_t or using such data-driven methods as AIC and cross-validation. See Fan, Yao and Cai [52] for further discussion on the choice of s_t. For exchange rate changes, Hong and Lee [85] choose s_t as the difference between the exchange rate at time $t-1$ and the moving average of the most recent L periods of exchange rates at time $t-1$. The moving average is a proxy for the local trend at time $t-1$. Intuitively, this choice of s_t is expected to reveal useful information on the direction of changes.

To justify the use of the functional coefficient model, CFY [24] suggest a goodness-of-fit test for an AR(d) model against a functional coefficient model. The null hypothesis of AR(d) can be stated as

$$\mathbb{H}_0 : a_j(s_t) = \beta_j, \quad j = 0, 1, \ldots, d\,,$$

where β_j is the autoregressive coefficient in AR(d). Under $\mathbb{H}_0$, $\{y_t\}$ is linear in mean conditional on X_t. Under the alternative to $\mathbb{H}_0$, the autoregressive coefficients depend on s_t and the AR(d) model suffers from "neglected nonlinearity". To test $\mathbb{H}_0$, CFY compares the residual sum of squares (RSS) under $\mathbb{H}_0$

$$RSS_0 \equiv \sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \sum_{t=1}^{n} \left[Y_t - \hat{\beta}_0 - \sum_{j=1}^{d} \hat{\beta}_j Y_{t-j} \right]^2$$

with the RSS under the alternative

$$RSS_1 \equiv \sum_{t=1}^{n} \tilde{\varepsilon}_t^2 = \sum_{t=1}^{n} \left[Y_t - \hat{a}_0(s_t) - \sum_{j=1}^{d} \hat{a}_j(s_t) Y_{t-j} \right]^2 .$$

The test statistic is $T_n = (RSS_0 - RSS_1)/RSS_1$. We reject $\mathbb{H}_0$ for large values of T_n. CFY suggest the following bootstrap method to obtain the p-value of T_n: (i) generate the bootstrap residuals $\{\varepsilon_t^b\}_{t=1}^n$ from the centered residuals $\tilde{\varepsilon}_t - \bar{\varepsilon}$ where $\bar{\varepsilon} \equiv n^{-1} \sum_{t=1}^n \tilde{\varepsilon}_t$ and define $y_t^b \equiv X_t' \hat{\beta} + \varepsilon_t^b$, where $\hat{\beta}$ is the OLS estimator for AR(d); (ii) calculate the bootstrap statistic T_n^b using the bootstrap sample $\{y_t^b, X_t', s_t\}_{t=1}^n$; (iii) repeat steps (i) and (ii) B times ($b = 1, \ldots, B$) and approximate the bootstrap p-value of T_n by $B^{-1} \sum_{b=1}^B \mathbf{1}(T_n^b \geq T_n)$. See Hong and Lee [85] for empirical application of the functional coefficient model to forecasting foreign exchange rates.

Nonparametric Regression

Let $\{y_t, x_t\}$, $t = 1, \ldots, n$, be stochastic processes, where y_t is a scalar and $x_t = (x_{t1}, \ldots, x_{tk})$ is a $1 \times k$ vector which may contain the lagged values of y_t. Consider the regression model

$$y_t = m(x_t) + u_t$$

where $m(x_t) = \mathbb{E}\left(y_t | x_t\right)$ is the true but unknown regression function and u_t is the error term such that $\mathbb{E}(u_t | x_t) = 0$.

If $m(x_t) = g(x_t, \delta)$ is a correctly specified family of parametric regression functions then $y_t = g(x_t, \delta) + u_t$ is a correct model and, in this case, one can construct a consistent least squares (LS) estimator of $m(x_t)$ given by $g(x_t, \hat{\delta})$, where $\hat{\delta}$ is the LS estimator of the parameter δ.

In general, if the parametric regression $g(x_t, \delta)$ is incorrect or the form of $m(x_t)$ is unknown then $g(x_t, \hat{\delta})$ may not be a consistent estimator of $m(x_t)$. For this case, an alternative approach to estimate the unknown $m(x_t)$ is to use the consistent nonparametric kernel regression estimator which is essentially a local constant LS (LCLS) estimator. To obtain this estimator take a Taylor series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= m(x) + e_t \end{aligned}$$

where $e_t = (x_t - x) m^{(1)}(x) + \frac{1}{2}(x_t - x)^2 m^{(2)}(x) + \cdots + u_t$ and $m^{(s)}(x)$ represents the sth derivative of $m(x)$ at $x_t = x$. The LCLS estimator can then be derived by minimizing

$$\sum_{t=1}^{n} e_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - m(x))^2 K_{tx}$$

with respect to constant $m(x)$, where $K_{tx} = K\left(\frac{x_t - x}{h}\right)$ is a decreasing function of the distances of the regressor vector x_t from the point $x = (x_1, \ldots, x_k)$, and $h \to 0$ as $n \to \infty$ is the window width (smoothing parameter) which determines how rapidly the weights decrease as the distance of x_t from x increases. The LCLS estimator so estimated is

$$\hat{m}(x) = \frac{\sum_{t=1}^{n} y_t K_{tx}}{\sum_{t=1}^{n} K_{tx}} = (\mathbf{i}' \mathbf{K}(x) \mathbf{i})^{-1}\, \mathbf{i}' \mathbf{K}(x) \mathbf{y}$$

where $\mathbf{K}(x)$ is the $n \times n$ diagonal matrix with the diagonal elements K_{tx} $(t = 1, \ldots, n)$, $\mathbf{i}$ is an $n \times 1$ column vector of unit elements, and $\mathbf{y}$ is an $n \times 1$ vector with elements y_t $(t = 1, \ldots, n)$. The estimator $\hat{m}(x)$ is due to Nadaraya [118] and Watson [155] (NW) who derived this in an alternative way. Generally $\hat{m}(x)$ is calculated at the data points x_t, in which case we can write the leave-one out estimator as

$$\hat{m}(x) = \frac{\sum_{t'=1,t'\neq t}^{n} y_{t'} K_{t't}}{\sum_{t'=1,t'\neq t}^{n} K_{t't}} ,$$

where $K_{t't} = K\frac{x_{t'}-x_t}{h}$. The assumption that $h \to 0$ as $n \to \infty$ gives $x_t - x = O(h) \to 0$ and hence $\mathbb{E}e_t \to 0$ as $n \to \infty$. Thus the estimator $\hat{m}(x)$ will be consistent under certain smoothing conditions on h, K, and $m(x)$. In small samples however $\mathbb{E}e_t \neq 0$ so $\hat{m}(x)$ will be a biased estimator, see [122] for details on asymptotic and small sample properties.

An estimator which has a better small sample bias and hence the mean square error (MSE) behavior is the local linear LS (LLLS) estimator. In the LLLS estimator we take a first order Taylor-Series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t = m(x) + (x_t - x)m^{(1)}(x) + v_t \\ &= \alpha(x) + x_t\beta(x) + v_t \\ &= X_t\delta(x) + v_t \end{aligned}$$

where $X_t = (1\ x_t)$ and $\delta(x) = [\alpha(x)\ \beta(x)']'$ with $\alpha(x) = m(x) - x\beta(x)$ and $\beta(x) = m^{(1)}(x)$. The LLLS estimator of $\delta(x)$ is then obtained by minimizing

$$\sum_{t=1}^{n} v_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - X_t\delta(x))^2 K_{tx}$$

sand it is given by

$$\tilde{\delta}(x) = (\mathbf{X}'\mathbf{K}(x)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(x)\mathbf{y} . \qquad (20)$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix with the tth row X_t ($t = 1, \ldots, n$).

The LLLS estimator of $\alpha(x)$ and $\beta(x)$ can be calculated as $\tilde{\alpha}(x) = (1\ 0)\tilde{\delta}(x)$ and $\tilde{\beta}(x) = (0\ 1)\tilde{\delta}(x)$. This gives

$$\tilde{m}(x) = (1\ x)\tilde{\delta}(x) = \tilde{\alpha}(x) + x\tilde{\beta}(x) .$$

Obviously when $X = \mathbf{i}$, $\tilde{\delta}(x)$ reduces to the NW's LCLS estimator $\hat{m}(x)$. An extension of the LLLS is the local polynomial LS (LPLS) estimators, see [50].

In fact one can obtain the local estimators of a general nonlinear model $g(x_t, \delta)$ by minimizing

$$\sum_{t=1}^{n} [y_t - g(x_t, \delta(x))]^2 K_{tx}$$

with respect to $\delta(x)$. For $g(x_t, \delta(x)) = X_t\delta(x)$ we get the LLLS in (20). Further when $h = \infty$, $K_{tx} = K(0)$ is a constant so that the minimization of $K(0)\sum[y_t - g(x_t, \delta(x))]^2$ is the same as the minimization of $\sum[y_t - g(x_t, \delta)]^2$, that is the local LS becomes the global LS estimator $\hat{\delta}$.

The LLLS estimator in (20) can also be interpreted as the estimator of the functional coefficient (varying coefficient) linear regression model

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= X_t\delta(x_t) + u_t \end{aligned}$$

where $\delta(x_t)$ is approximated locally by a constant $\delta(x_t) \simeq \delta(x)$. The minimization of $\sum u_t^2 K_{tx}$ with respect to $\delta(x)$ then gives the LLLS estimator in (20), which can be interpreted as the LC varying coefficient estimator. An extension of this is to consider the linear approximation $\delta(x_t) \simeq \delta(x) + D(x)(x_t - x)'$ where $D(x) = \frac{\partial\delta(x_t)}{\partial x_t'}$ evaluated at $x_t = x$. In this case

$$\begin{aligned} y_t &= m(x_t) + u_t = X_t\delta(x_t) + u_t \\ &\simeq X_t\delta(x) + X_t D(x)(x_t - x)' + u_t \\ &= X_t\delta(x) + [(x_t - x) \otimes X_t]vecD(x) + u_t \\ &= X_t^x\delta^x(x) + u_t \end{aligned}$$

where $X_t^x = [X_t \quad (x_t - x) \otimes X_t]$ and $\delta^x(x) = [\delta(x)'\ (vecD(x))']'$. The LL varying coefficient estimator of $\delta^x(x)$ can then be obtained by minimizing

$$\sum_{t=1}^{n} [y_t - X_t^x\delta^x(x)]^2 K_{tx}$$

with respect to $\delta^x(x)$ as

$$\dot{\delta}^x(x) = (\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{X}^x)^{-1}\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{y} . \qquad (21)$$

From this $\dot{\delta}(x) = (\mathbf{I}\ 0)\dot{\delta}^x(x)$, and hence

$$\dot{m}(x) = (1\ x\ 0)\dot{\delta}^x(x) = (1\ x)\dot{\delta}(x) .$$

The above idea can be extended to the situations where $\xi_t = (x_t\ z_t)$ such that

$$\mathbb{E}(y_t|\xi_t) = m(\xi_t) = m(x_t, z_t) = X_t\delta(z_t) ,$$

where the coefficients are varying with respect to only a subset of ξ_t; z_t is $1 \times l$ and ξ_t is $1 \times p$, $p = k + l$. Examples of these include functional coefficient autoregressive models of Chen and Tsay [26] and CFY [24], random coefficient models of Raj and Ullah [128], smooth transition autoregressive models of Granger and Teräsvirta [72], and threshold autoregressive models of Tong [149].

To estimate $\delta(z_t)$ we can again do a local constant approximation $\delta(z_t) \simeq \delta(z)$ and then minimize $\sum[y_t - X_t\delta(z)]^2 K_{tz}$ with respect to $\delta(z)$, where

$K_{tz} = K(\frac{z_t - z}{h})$. This gives the LC varying coefficient estimator

$$\tilde{\delta}(z) = (\mathbf{X}'\mathbf{K}(z)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(z)\mathbf{y} \tag{22}$$

where $\mathbf{K}(z)$ is a diagonal matrix of $K_{tz}, t = 1, \ldots, n$. When $z = x$, (22) reduces to the LLLS estimator $\tilde{\delta}(x)$ in (20).

CFY [24] consider a local linear approximation $\delta(z_t) \simeq \delta(z) + D(z)(z_t - z)'$. The LL varying coefficient estimator of CFY is then obtained by minimizing

$$\begin{aligned}
&\sum_{t=1}^{n}[y_t - X_t\delta(z_t)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n}[y_t - X_t\delta(z) - [(z_t - z) \otimes X_t]vecD(z)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n}[y_t - X_t^z\delta^z(z)]^2 K_{tz}
\end{aligned}$$

with respect to $\delta^z(z) = [\delta(z)' \quad (vecD(z))']'$ where $X_t^z = [X_t \quad (z_t - z) \otimes X_t]$. This gives

$$\ddot{\delta}^z(z) = (\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{X}^z)^{-1}\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{y}\,, \tag{23}$$

and $\ddot{\delta}(z) = (\mathbf{I}\ 0)\ddot{\delta}^z(z)$. Hence

$$\ddot{m}(\xi) = (1\ x\ 0)\ddot{\delta}^z(z) = (1\ x)\ddot{\delta}(z)\,.$$

For the asymptotic properties of these varying coefficient estimators, see CFY [24]. When $z = x$, (23) reduces to the LL varying coefficient estimator $\dot{\delta}^x(x)$ in (21). See Lee and Ullah [98] for more discussion of these models and issues of testing nonlinearity.

Regime Switching Autoregressive Model Between Unit Root and Stationary Root

To avoid the usual dichotomy between unit-root nonstationarity and stationarity, we may consider models that permit two regimes of unit root nonstationarity and stationarity.

One model is the Innovation Regime-Switching (IRS) model of Kuan, Huang, and Tsay [96]. Intuitively, it may be implausible to believe that all random shocks exert only one effect (permanent or transitory) on future financial asset prices in a long time span. This intuition underpins the models that allow for breaks, stochastic unit root, or regime switching. As an alternative, Kuan, Huang, and Tsay [96] propose the IRS model that permits the random shock in each period to be permanent or transitory, depending on a switching mechanism, and hence admits distinct dynamics (unit-root nonstationarity or stationarity) in different periods. Under the IRS framework, standard unit-root models and stationarity models are just two extreme cases. By applying the IRS model to real exchange rate, they circumvent the difficulties arising from unit-root (or stationarity) testing. They allow the data to speak for themselves, rather than putting them in the straitjacket of unit-root nonstationarity or stationarity. Huang and Kuan [90] re-examine long-run PPP based on the IRS model and their empirical study on US/UK real exchange rates shows that there are both temporary and permanent influences on the real exchange rate such that approximately 42% of the shocks in the long run are more likely to have a permanent effect. They also found that transitory shocks dominate in the fixed-rate regimes, yet permanent shocks play a more important role during the floating regimes. Thus, the long-run PPP is rejected due to the presence of a significant amount of permanent shocks, but there are still long periods of time in which the deviations from long-run PPP are only transitory.

Another model is a threshold unit root (TUR) model or threshold integrated moving average (TIMA) model of Gonzalo and Martíneza [65]. Based on this model they examine whether large and small shocks have different long-run effects, as well as whether one of them is purely transitory. They develop a new nonlinear permanent – transitory decomposition, that is applied to US stock prices to analyze the quality of the stock market.

Comparison of these two models with the linear autoregressive model with a unit root or a stationary AR model for the out-of-sample forecasting remains to be examined empirically.

Bagging Nonlinear Forecasts

To improve on unstable forecasts, bootstrap aggregating or bagging is introduced by Breiman [19]. Lee and Yang [100] show how bagging works for binary and quantile predictions. Lee and Yang [100] attributed part of the success of the bagging predictors to the small sample estimation uncertainties. Therefore, a question that may arise is that whether the good performance of bagging predictors critically depends on algorithms we employ in nonlinear estimation.

They find that bagging improves the forecasting performance of predictors on highly nonlinear regression models – e. g., artificial neural network models, especially when the sample size is limited. It is usually hard to choose the number of hidden nodes and the number of inputs (or lags), and to estimate the large number of parameters in an ANN model. Therefore, a neural network model generate poor predictions in a small sample. In such cases, bagging can do a valuable job to improve the forecasting performance as shown in [100], confirming the result of Breiman [20]. A bagging predictor is a combined predictor formed over a set of training sets to smooth out the "instability" caused by parameter estimation uncertainty and model uncertainty. A predictor is said to be "unstable" if a small change in the training set will lead to a significant change in the predictor [20].

As bagging would be valuable in nonlinear forecasting, in this section, we will show how a bagging predictor may improve the predicting performance of its underlying predictor. Let

$$\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t} \quad (t = R, \ldots, T)$$

be a training set at time t and let $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ be a forecast of Y_{t+1} or of the binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$ using this training set $\mathcal{D}_t$ and the explanatory variable vector $\mathbf{X}_t$. The optimal forecast $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ for Y_{t+1} will be the conditional mean of Y_{t+1} given $\mathbf{X}_t$ under the squared error loss function, or the conditional quantile of Y_{t+1} on $\mathbf{X}_t$ if the loss is a tick function. Below we also consider the binary forecast for $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$.

Suppose each training set $\mathcal{D}_t$ consists of R observations generated from the underlying probability distribution $\mathbf{P}$. The forecast $\{\varphi(\mathbf{X}_t, \mathcal{D}_t)\}_{t=R}^{T}$ can be improved if more training sets were able to be generated from $\mathbf{P}$ and the forecast can be formed from averaging the multiple forecasts obtained from the multiple training sets. Ideally, if $\mathbf{P}$ were known and multiple training sets $\mathcal{D}_t^{(j)}$ $(j = 1, \ldots, J)$ may be drawn from $\mathbf{P}$, an ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ can be constructed by the weighted averaging of $\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)})$ over j, i. e.,

$$\varphi_A(\mathbf{X}_t) \equiv \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t) \equiv \sum_{j=1}^{J} w_{j,t}\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)}) \,,$$

where $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ denotes the expectation over $\mathbf{P}$, $w_{j,t}$ is the weight function with $\sum_{j=1}^{J} w_{j,t} = 1$, and the subscript A in φ_A denotes "aggregation".

Lee and Yang [100] show that the ensemble aggregating predictor $\varphi_A(X_t)$ has not a larger expected loss than the original predictor $\varphi(X_t, \mathcal{D}_t)$. For any convex loss function $c(\cdot)$ on the forecast error z_{t+1}, we will have

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) \geq \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1})),$$

where $\mathbb{E}_{\mathcal{D}_t}(z_{t+1})$ is the aggregating forecast error, and $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}\{\mathbb{E}_{\mathcal{D}_t}(\cdot)\,|X_t\}]$ denotes the expectation $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ taken over $\mathbf{P}$ (i. e., averaging over the multiple training sets generated from $\mathbf{P}$), then taking an expectation of Y_{t+1} conditioning on X_t, and then taking an expectation of X_t. Similarly we define the notation $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}(\cdot)\,|X_t]$. Therefore, the aggregating predictor will always have no larger expected cost than the original predictor for a convex loss function $\varphi(X_t, D_t)$. The examples of the convex loss function includes the squared error loss and a tick (or check) loss $\rho_\alpha(z) \equiv [\alpha - \mathbf{1}(z < 0)]z$.

How much this aggregating predictor can improve depends on the distance between $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1})$ and $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. We can define this distance by $\Delta \equiv \mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) - \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. Therefore, the effectiveness of the aggregating predictor depends on the *convexity* of the cost function. The more convex is the cost function, the more effective this aggregating predictor can be. If the loss function is the squared error loss, then it can be shown that $\Delta = \mathbb{V}_{\mathcal{D}_t}\left[\varphi(\mathbf{X}_t, \mathcal{D}_t)\right]$ is the variance of the predictor, which measures the "instability" of the predictor. See Lee and Yang [100], Proposition 1, and Breiman [20]. If the loss is the tick function, the effectiveness of bagging is also different for different quantile predictions: bagging works better for tail-quantile predictions than for mid-quantile predictions.

In practice, however, $\mathbf{P}$ is not known. In that case we may estimate $\mathbf{P}$ by its empirical distribution, $\hat{\mathbf{P}}(\mathcal{D}_t)$, for a given $\mathcal{D}_t$. Then, from the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_t)$, multiple training sets may be drawn by the bootstrap method. Bagging predictors, $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$, can then be computed by taking weighted average of the predictors trained over a set of bootstrap training sets. More specifically, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be obtained in the following steps:

1. Given a training set of data at time t, $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t}$, construct the jth bootstrap sample $\mathcal{D}_t^{*(j)} \equiv \{(Y_s^{*(j)}, \mathbf{X}_{s-1}^{*(j)})\}_{s=t-R+1}^{t}$, $j = 1, \ldots, J$, according to the empirical distribution of $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$.
2. Train the model (estimate parameters) from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
3. Compute the bootstrap predictor $\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})$ from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
4. Finally, for mean and quantile forecast, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by averaging over J bootstrap predictors

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \sum_{j=1}^{J} \hat{w}_{j,t}\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})\,;$$

and for binary forecast, the bagging binary predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by majority voting over J bootstrap predictors:

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \mathbf{1}\left(\sum_{j=1}^{J} \hat{w}_{j,t}\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) > 1/2\right)$$

with $\sum_{j=1}^{J} \hat{w}_{j,t} = 1$ in both cases.

One concern of applying bagging to time series is whether a bootstrap can provide a sound simulation sample for dependent data, for which the bootstrap is required to be consistent. It has been shown that some bootstrap procedure (such as moving block bootstrap) can provide consistent densities for moment estimators and quantile estimators. See, e. g., Fitzenberger [54].

Nonlinear Forecasting Models for the Conditional Variance

Nonlinear Parametric Models for Volatility

Volatility models are of paramount importance in financial economics. Issues such as portfolio allocation, option pricing, risk management, and generally any decision making under uncertainty rely on the understanding and forecasting of volatility. This is one of the most active ares of research in time series econometrics. Important surveys as in Bollerslev, Chou, and Kroner [15], Bera and Higgins [13], Bollerslev, Engle, and Nelson [16], Poon and Granger [125], and Bauwens, Laurent, and Rombouts [12] attest to the variety of issues in volatility research. The motivation for the introduction of the first generation of volatility models namely the ARCH models of Engle [44] was to account for clusters of activity and fat-tail behavior of financial data. Subsequent models accounted for more complex issues. Among others and without being exclusive, we should mention issues related to asymmetric responses of volatility to news, probability distribution of the standardized innovations, i.i.d. behavior of the standardized innovation, persistence of the volatility process, linkages with continuous time models, intraday data and unevenly spaced observations, seasonality and noise in intraday data. The consequence of this research agenda has been a vast array of specifications for the volatility process.

Suppose that the return series $\{y_t\}_{t=1}^{T+1}$ of a financial asset follows the stochastic process $y_{t+1} = \mu_{t+1} + \varepsilon_{t+1}$, where $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_{t+1}(\theta)$ and $\mathbb{E}(\varepsilon_{t+1}^2|\mathcal{F}_t) = \sigma_{t+1}^2(\theta)$ given the information set $\mathcal{F}_t$ (σ-field) at time t. Let $z_{t+1} \equiv \varepsilon_{t+1}/\sigma_{t+1}$ have the conditional normal distribution with zero conditional mean and unit conditional variance. Volatility models can be classified in three categories: MA family, ARCH family, and stochastic volatility (SV) family.

The simplest method to forecast volatility is to calculate a historical moving average variance, denoted as MA(m), or an exponential weighted moving average (EWMA):

MA(m)	$\sigma_t^2 = \frac{1}{m}\sum_{j=1}^{m}(y_{t-j} - \hat{\mu}_t^m)^2, \quad \hat{\mu}_t^m = \frac{1}{m}\sum_{j=1}^{m} y_{t-j}$
EWMA	$\sigma_t^2 = (1-\lambda)\sum_{j=1}^{t-1}\lambda^{j-1}(y_{t-j} - \hat{\mu}_t)^2,$ $\hat{\mu}_t = \frac{1}{t-1}\sum_{j=1}^{t-1} y_{t-j}$

In the EWMA specification, a common practice is to fix the λ parameter, for instance $\lambda = 0.94$ [129]. For these two MA family models, there are not parameters to estimate.

Second, the ARCH family is very extensive with many variations on the original model ARCH(p) of Engle [44]. Some representative models are: GARCH model of Bollerslev [14]; Threshold GARCH (T-GARCH) of Glosten et al. [60]; Exponential GARCH (E-GARCH) of Nelson [120]; quadratic GARCH models (Q-GARCH) as in Sentana [135]; Absolute GARCH (ABS-GARCH) of Taylor [143] and Schwert [134] and Smooth Transition GARCH (ST-GARCH) of González-Rivera [61].

ARCH(p)	$\sigma_t^2 = \omega + \sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2$
GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2$
I-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2,\ \alpha + \beta = 1$
T-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2\mathbf{1}(\varepsilon_{t-1} \geq 0)$
ST-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 F(\varepsilon_{t-1}, \delta)$ with $F(\varepsilon_{t-1}, \delta) = [1 + \exp(\delta\varepsilon_{t-1})]^{-1} - 0.5$
E-GARCH	$\ln\sigma_t^2 = \omega + \beta\ln\sigma_{t-1}^2 + \alpha[\lvert z_{t-1}\rvert - cz_{t-1}]$
Q-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(\varepsilon_{t-1} + \gamma)^2$
ABS-GARCH	$\sigma_t = \omega + \beta\sigma_{t-1} + \alpha\lvert\varepsilon_{t-1}\rvert$

The EWMA specification can be viewed as an integrated GARCH model with $\omega = 0, \alpha = \lambda$, and $\beta = 1-\lambda$. In the T-GARCH model, the parameter γ allows for possible asymmetric effects of positive and negative innovations. In Q-GARCH models, the parameter γ measures the extent of the asymmetry in the news impact curve. For the ST-GARCH model, the parameter γ measures the asymmetric effect of positive and negative shocks, and the parameter $\delta > 0$ measures the smoothness of the transition between regimes, with a higher value of δ making ST-GARCH closer to T-GARCH.

Third, the stationary SV model of Taylor [143] with η_t is i.i.d. N $(0, \sigma_\eta^2)$ and ξ_t is i.i.d. N$(0, \pi^2/2)$ is a representative member of the SV family.

SV	$\sigma_t^2 = \exp(0.5h_t), \quad \ln(y_t^2) = -1.27 + h_t + \xi_t,$ $h_t = \gamma + \phi h_{t-1} + \eta_t.$

With so many models, the natural question becomes which one to choose. There is not a universal answer to this question. The best model depends upon the objectives of the user. Thus, given an objective function, we search for the model(s) with the best predictive ability controlling for possible biases due to “data snooping” [105]. To compare the relative performance of volatility models, it is customary to choose either a statistical loss function or an economic loss function.

The preferred statistical loss functions are based on moments of forecast errors (mean-error, mean-squared error, mean absolute error, etc.). The best model will minimize a function of the forecast errors. The volatility forecast is often compared to a measure of realized volatility. With financial data, the common practice has been to take squared returns as a measure of realized volatility. However, this practice is questionable. Andersen and Bollerslev [2] argued that this measure is a noisy estimate, and proposed the use of the intra-day (at each five minutes interval) squared returns to calculate the daily realized volatility. This measure requires intra-day data, which is subject to the variation introduced by the bid-ask spread and the irregular spacing of the price quotes.

Some other authors have evaluated the performance of volatility models with criteria based on economic loss functions. For example, West, Edison, and Cho [157] considered the problem of portfolio allocation based on models that maximize the utility function of the investor. Engle, Kane, and Noh [46] and Noh, Engle, and Kane [121] considered different volatility forecasts to maximize the trading profits in buying/selling options. Lopez [107] considered probability scoring rules that were tailored to a forecast user’s decision problem and confirmed that the choice of loss function directly affected the forecast evaluation of different models. Brooks and Persand [21] evaluated volatility forecasting in a financial risk management setting in terms of Value-at-Risk (VaR). The common feature to these branches of the volatility literature is that none of these has controlled for forecast dependence across models and the inherent biases due to data-snooping.

Controlling for model dependence [160], González-Rivera, Lee, and Mishra [62] evaluate fifteen volatility models for the daily returns to the SP500 index according to their out-of-sample forecasting ability. The forecast evaluation is based, among others, on two economic loss functions: an option pricing formula and a utility function; and a statistical loss function: a goodness-of-fit based on a Value-at-Risk (VaR) calculation. For option pricing, volatility is the only component that is not observable and it needs to be estimated. The loss function assess the difference between the actual price of a call option and the estimated price, which is a function of the estimated volatility of the stock. The second economic loss function refers to the problem of wealth allocation. An investor wishes to maximize her utility allocating wealth between a risky asset and a risk-free asset. The loss function assesses the performance of the volatility estimates according to the level of utility they generate. The statistical function based on the goodness-of-fit of a VaR calculation is important for risk management. The main objective of VaR is to calculate extreme losses within a given probability of occurrence, and the estimation of the volatility is central to the VaR measure. The preferred models depend very strongly upon the loss function chosen by the user. González-Rivera, Lee, and Mishra [62] find that, for option pricing, simple models such as the exponential weighted moving average (EWMA) proposed by Riskmetrics [64] performed as well as any GARCH model. For an utility loss function, an asymmetric quadratic GARCH model is the most preferred. For VaR calculations, a stochastic volatility model dominates all other models.

Nonparametric Models for Volatility

Ziegelmann [163] considers the kernel smoothing techniques that free the traditional parametric volatility estimators from the constraints related to their specific models. He applies the nonparametric local ‘exponential’ estimator to estimate conditional volatility functions, ensuring its nonnegativity. Its asymptotic properties are established and compared with those for the local linear estimator for the volatility model of Fan and Yao [51]. Long, Su, and Ullah [106] extend this idea to semiparametric multivariate GARCH and show that there may exist substantial out-of-sample forecasting gain over the parametric models. This gain accounts for the presence of nonlinearity in the conditional variance-covariance that is neglected in parametric linear models.

Forecasting Volatility Using High Frequency Data

Using high-frequency data, quadratic variation may be estimated using realized volatility (RV). Andersen, Bollerslev, Diebold, and Labys [3] and Barndorff-Nielsen and Shephard [11] establish that RV, defined as the sum of squared intraday returns of small intervals, is an asymptotically unbiased estimator of the unobserved quadratic variation as the interval length approaches zero. Besides the use of high frequency information in volatility estimation, volatility forecasting using high frequency information has been addressed as well. In an application to volatility prediction, Ghysels, Santa-Clara, and Valkanov [58] investigate the predictive power of various regressors (lagged realized volatility, squared return, realized power, and daily range) for future volatility forecasting. They find that the best predictor is realized power (sum of absolute intraday returns), and more interestingly, direct use of intraday squared returns in mixed data sampling (MIDAS) regressions does not necessarily lead to better volatility forecasts.

Andersen, Bollerslev, Diebold, and Labys [4] represent another approach to forecasting volatility using RV. The model they propose is a fractional integrated AR model: ARFI(5, d) for logarithmic RV's obtained from foreign exchange rates data of 30-minute frequency and demonstrate the superior predictive power of their model.

Alternatively, Corsi [32] proposes the heterogeneous autoregressive (HAR) model of RV, which is able to reproduce long memory. McAleer and Medeiros [115] propose a new model that is a multiple regime smooth transition (ST) extension of the HAR model, which is specifically designed to model the behavior of the volatility inherent in financial time series. The model is able to describe simultaneously long memory as well as sign and size asymmetries. They apply the model to several Dow Jones Industrial Average index stocks using transaction level data from the Trades and Quotes database that covers ten years of data, and find strong support for long memory and both sign and size asymmetries. Furthermore, they show that the multiple regime smooth transition HAR model, when combined with the linear HAR model, is flexible for the purpose of forecasting volatility.

Forecasting Beyond Mean and Variance

In the previous section, we have surveyed the major developments in nonlinear time series, mainly modeling the conditional mean and the conditional variance of financial returns. However it is not clear yet that any of those nonlinear models may generate profits after accounting for various market frictions and transactions costs. Therefore, some research efforts have been directed to investigate other aspects of the conditional density of returns such as higher moments, quantiles, directions, intervals, and the density itself. In this section, we provide a brief survey on forecasting these other features.

Forecasting Quantiles

The optimal forecast of a time series model depends on the specification of the loss function. A symmetric quadratic loss function is the most prevalent in applications due to its simplicity. Under symmetric quadratic loss, the optimal forecast is simply the conditional mean. An asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67].

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \dots, Z_t\}$ of size R or the whole past sample $\{Z_1, \dots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^{T}$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f^*_{t,1}$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) ,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e^*_{t+1} = Y_{t+1} - f^*_{t,1} .$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f^*_{t,1}) \mathrm{d}F_t(y) = 0 .$$

When we interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f^*_{t,1}) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1}) | I_t \right)$$

Based on the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1})$, the condition for forecast optimality is:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t\right) = 0 \quad a.s. ,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

To see this we consider the following two examples. First, when the loss function is the squared error loss

$$c(Y_{t+1} - f_{t,1}) = (Y_{t+1} - f_{t,1})^2 ,$$

the generalized forecast error will be $g_{t+1} \equiv \frac{\partial}{\partial f_t} c(Y_{t+1} - f^*_{t,1}) = -2e^*_{t+1}$ and thus $\mathbb{E}\left(e^*_{t+1}|I_t\right) = 0$ *a.s.*, which implies that the optimal forecast

$$f^*_{t,1} = \mathbb{E}\left(Y_{t+1}|I_t\right)$$

is the conditional mean. Next, when the loss is the check function, $c(e) = \left[\alpha - \mathbf{1}(e < 0)\right] \cdot e \equiv \rho_\alpha(e_{t+1})$, the optimal forecast $f_{t,1}$, for given $\alpha \in (0, 1)$, minimizing

$$\min_{f_{t,1}} \mathbb{E}\left[c(Y_{t+1} - f_{t,1})|I_t\right]$$

can be shown to satisfy

$$\mathbb{E}\left[\alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})|I_t\right] = 0 \quad a.s.$$

Hence, $g_{t+1} \equiv \alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})$ is the generalized forecast error. Therefore,

$$\alpha = \mathbb{E}\left[\mathbf{1}(Y_{t+1} < f^*_{t,1})|I_t\right] = \Pr(Y_{t+1} \leq f^*_{t,1}|I_t)\,,$$

and the optimal forecast $f^*_{t,1} = q^\alpha\left(Y_{t+1}|I_t\right) \equiv q^\alpha_t$ is the conditional α-quantile.

Forecasting conditional quantiles are of paramount importance for risk management, which nowdays is key activity in financial institutions due to the increasing financial fragility in emerging markets and the extensive use of derivative products over the last decade. A risk measurement methodology called Value-at-Risk (VaR) has received a great attention from both regulatory and academic fronts. During a short span of time, numerous papers have studied various aspects of the VaR methodology. Bao, Lee, and Saltoglu [8] examine the relative out-of-sample predictive performance of various VaR models.

An interesting VaR model is the CaViaR (conditional autoregressive Value-at-Risk) model suggested by Engle and Manganelli [47]. They estimate the VaR from a quantile regression rather than inverting a conditional distribution. The idea is similar to the GARCH modeling in that VaR is modeled autoregressively

$$q_t(\alpha) = a_0 + a_1 q_{t-1}(\alpha) + h(x_t|\theta)\,,$$

where $x_t \in \mathcal{F}_{t-1}$, θ is a parameter vector, and $h(\cdot)$ is a function to explain the VaR model. Depending on the specification of $h(\cdot)$, the CaViaR model may be

$$q_t(\alpha) = a_0 + a_1 q_{t-1}(\alpha) + a_2|r_{t-1}|\,,$$

$$q_t(\alpha) = a_0 + a_1 q_{t-1}(\alpha) + a_2|r_{t-1}| + a_3|r_{t-1}| \cdot \mathbf{1}(r_{t-1} < 0),$$

where the second model allow nonlinearity (asymmetry) similarly to the asymmetric GARCH models.

Bao, Lee, and Saltoglu [8] compare various VaR models. Their results show that the CaViaR quantile regression models of Engle and Manganelli [47] have shown some success in predicting the VaR risk measure for various periods of time, and it is generally more stable than the models that invert a distribution function.

Forecasting Directions

It is well known that, while financial returns $\{Y_t\}$ may not be predictable, their variance, sign, and quantiles may be predictable. Christofferson and Diebold [27] show that binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$, where $\mathbf{1}(\cdot)$ takes the value of 1 if the statement in the parenthesis is true, and 0 otherwise, is predictable when some conditional moments are time varying, Hong and Lee [86], Hong and Chung [85], Linton and Whang [104], Lee and Yang [100] among many others find some evidence that the directions of stock returns and foreign exchange rate changes are predictable.

Lee and Yang [100] also show that forecasting quantiles and forecasting binary (directional) forecasts are related, in that the former may lead to the latter. As noted by Powell [126], using the fact that for any monotonic function $h(\cdot)$, $q^\alpha_t(h(Y_{t+1})|\mathbf{X}_t) = h(q^\alpha_t(Y_{t+1}|\mathbf{X}_t))$, which follows immediately from observing that $\Pr(Y_{t+1} < y|\mathbf{X}_t) = \Pr[h(Y_{t+1}) < h(y)|\mathbf{X}_t]$, and noting that the indicator function is monotonic, $q^\alpha_t(G_{t+1}|\mathbf{X}_t) = q^\alpha_t(\mathbf{1}(Y_{t+1} > 0)|\mathbf{X}_t) = \mathbf{1}(q^\alpha_t(Y_{t+1}|\mathbf{X}_t) > 0)$. Therefore, predictability of conditional quantiles of financial returns may imply predictability of conditional direction.

Probability Forecasts

Diebold and Rudebush [38] consider the probability forecasts for the turning points of the business cycle. They measure the accuracy of predicted probabilities, that is the average distance between the predicted probabilities and observed realization (as measured by a zero-one dummy variable). Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time $t (R \leq t < T + 1)$, to estimate a model. We then have time series of $P = T - R + 1$ probability forecasts $\{p_{t+1}\}^T_{t=R}$ where p_t is the predicted probability of the occurrence of an event (e. g., business cycle turning point) in the next period $t + 1$. Let $\{d_{t+1}\}^T_{t=R}$ be the corresponding realization with $d_t = 1$ if a business cycle turning point (or any defined event) occurs in period t and $d_t = 0$ otherwise. The loss function analogous to the squared error is the Brier's score based on quadratic probability score (QPS):

$$QPS = P^{-1} \sum_{t=R}^{T} 2(p_t - d_t)^2\,.$$

The QPS ranges from 0 to 2, with 0 for perfect accuracy. As noted by Diebold and Rudebush [38], the use of the symmetric loss function may not be appropriate as a forecaster may be penalized more heavily for missing a call (making a type II error) than for signaling a false alarm (making a type I error). Another loss function is given by the log probability score (LPS)

$$LPS = -P^{-1} \sum_{t=R}^{T} \ln \left(p_t^{d_t} (1 - p_t)^{(1-d_t)} \right) ,$$

which is similar to the loss for the interval forecast. A large mistake is penalized more heavily under LPS than under QPS. More loss functions are discussed in Diebold and Rudebush [38].

Another loss function useful in this context is the Kuipers score (KS), which is defined by

$$KS = \text{Hit Rate} - \text{False Alarm Rate} ,$$

where Hit Rate is the fraction of the bad events that were correctly predicted as good events (power, or 1− probability of type II error), and False Alarm Rate is the fraction of good events that had been incorrectly predicted as bad events (probability of type I error).

Forecasting Interval

Suppose Y_t is a stationary series. Let the one-period ahead conditional interval forecast made at time t from a model be denoted as

$$J_{t,1}(\alpha) = (L_{t,1}(\alpha), U_{t,1}(\alpha)), \quad t = R, \ldots, T ,$$

where $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ are the lower and upper limits of the ex ante interval forecast for time $t+1$ made at time t with the coverage probability α. Define the indicator variable $X_{t+1}(\alpha) = \mathbf{1}[Y_{t+1} \in J_{t,1}(\alpha)]$. The sequence $\{X_{t+1}(\alpha)\}_{t=R}^{T}$ is i.i.d. Bernoulli (α). The optimal interval forecast would satisfy $\mathbb{E}(X_{t+1}(\alpha)|I_t) = \alpha$, so that $\{X_{t+1}(\alpha) - \alpha\}$ will be an MD. A better model has a larger expected Bernoulli log-likelihood

$$\mathbb{E}\alpha^{X_{t+1}(\alpha)}(1-\alpha)^{[1-X_{t+1}(\alpha)]} .$$

Hence, we can choose a model for interval forecasts with the largest out-of-sample mean of the predictive log-likelihood, which is defined by

$$P^{-1} \sum_{t=R}^{T} \ln \left(\alpha^{x_{t+1}(\alpha)} (1-\alpha)^{[1-x_{t+1}(\alpha)]} \right) .$$

Evaluation of Nonlinear Forecasts

In order to evaluate the possible superior predictive ability of nonlinear models, we need to compare competing models in terms of a certain loss function. The literature has recently been exploding on this issue. Examples are Granger and Newbold [69], Diebold and Mariano [37], West [156], White [160], Hansen [81], Romano and Wolf [130], Giacomini and White [59], etc. In different perspective, to test the optimality of a given model, Patton and Timmermann [123] examine various testable properties that should hold for an optimal forecast.

Loss Functions

The loss function (or cost function) is a crucial ingredient for the evaluation of nonlinear forecasts. When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c(Y_{t+h}, f_{t,h})$. The loss function can depend on the time of prediction and so it can be $c_{t+h}(Y_{t+h}, f_{t,h})$. If the loss function is not changing with time and does not depend on the value of the variable Y_{t+h}, the loss can be written simply as a function of the error only, $c_{t+h}(Y_{t+h}, f_{t,h}) = c(e_{t+h})$.

Granger [67] discusses the following required properties for a loss function: (i) $c(0) = 0$ (no error and no loss), (ii) $\min_e c(e) = 0$, so that $c(e) \geq 0$, and (iii) $c(e)$ is monotonically nondecreasing as e moves away from zero so that $c(e_1) \geq c(e_2)$ if $e_1 > e_2 > 0$ and if $e_1 < e_2 < 0$.

When $c_1(e), c_2(e)$ are both loss functions, Granger [67] shows that further examples of loss functions can be generated: $c(e) = ac_1(e) + bc_2(e)$, $a \geq 0, b \geq 0$ will be a loss function. $c(e) = c_1(e)^a c_2(e)^b$, $a > 0, b > 0$ will be a loss function. $c(e) = 1(e > 0)c_1(e) + 1(e < 0)c_2(e)$ will be a loss function. If $h(\cdot)$ is a positive monotonic nondecreasing function with $h(0)$ finite, then $c(e) = h(c_1(e)) - h(0)$ is a loss function.

Granger [68] notes that an expected loss (a risk measure) of financial return Y_{t+1} that has a conditional predictive distribution $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$ with $\mathbf{X}_t \in I_t$ may be written as

$$\mathbb{E}c(e) = A_1 \int_0^{\infty} |y - f|^p dF_t(y) + A_2 \int_{-\infty}^{0} |y - f|^p \mathrm{d}F_t(y),$$

with A_1, A_2 both > 0 and some $\theta > 0$. Considering the symmetric case $A_1 = A_2$, one has a class of volatility measures $V_p = \mathbb{E}\left[|y - f|^p\right]$, which includes the variance with $p = 2$, and mean absolute deviation with $p = 1$.

Ding, Granger, and Engle [39] study the time series and distributional properties of these measures empirically and show that the absolute deviations are found to have some particular properties such as the longest memory. Granger remarks that given that the financial returns are known to come from a long tail distribution, $p = 1$ may be more preferable.

Another problem raised by Granger is how to choose optimal L_p-norm in empirical works, to minimize $\mathbb{E}[|\varepsilon_t|^p]$ for some p to estimate the regression model $Y_t = \mathbb{E}(Y_t|X_t;\beta) + \varepsilon_t$. As the asymptotic covariance matrix of $\hat{\beta}$ depends on p, the most appropriate value of p can be chosen to minimize the covariance matrix. In particular, Granger [68] refers to a trio of papers [84,116,117] who find that the optimal $p = 1$ from Laplace and Cauchy distribution, $p = 2$ for Gaussian and $p = \infty$ (min/max estimator) for a rectangular distribution. Granger [68] also notes that in terms of the kurtosis κ, Harter [84] suggests to use $p = 1$ for $\kappa > 3.8$; $p = 2$ for $2.2 \leq \kappa \leq 3.8$; and $p = 3$ for $\kappa < 2.2$. In finance, the kurtosis of returns can be thought of as being well over 4 and so $p = 1$ is preferred.

Forecast Optimality

Optimal forecast of a time series model extensively depends on the specification of the loss function. Symmetric quadratic loss function is the most prevalent in applications due to its simplicity. The optimal forecast under quadratic loss is simply the conditional mean, but an asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67], as the expected loss function if formulated with the expectation taken with respect to the conditional distribution. Specification of the loss function defines the model under consideration.

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \ldots, Z_t\}$ of size R or the whole past sample $\{Z_1, \ldots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^{T}$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^*$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) ,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e_{t+1}^* = Y_{t+1} - f_{t,1.}^*$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}^*) \mathrm{d}F_t(y) = 0 .$$

When we may interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f_{t,1}^*) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*)|I_t\right)$$

the “generalized forecast error”, $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*)$, forms the condition of forecast optimality:

$$H_0 : \mathbb{E}\left(g_{t+1}|I_t\right) = 0 \quad a.s.,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

Forecast Evaluation of Nonlinear Transformations

Granger [67] note that it is implausible to use the same loss function for forecasting Y_{t+h} and for forecasting $h_{t+1} = h(Y_{t+h})$ where $h(\cdot)$ is some function, such as the log or the square, if one is interested in forecasting volatility. Suppose the loss functions $c_1(\cdot), c_2(\cdot)$ are used for forecasting Y_{t+h} and for forecasting $h(Y_{t+h})$, respectively. Let $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c_1(e_{t+1})$, for which the optimal forecast $f_{t,1}^*$ will be chosen from $\min_{f_{t,1}} \int_{-\infty}^{\infty} c_1(y - f_{t,1}) \mathrm{d}F_t(y)$, where $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$. Let $\varepsilon_{t+1} \equiv h_{t+1} - h_{t,1}$ will result in a cost of $c_2(\varepsilon_{t+1})$, for which the optimal forecast $h_{t,1}^*$ will be chosen from $\min_{h_{t,1}} \int_{-\infty}^{\infty} c_2(h - h_{t,1}) \mathrm{d}H_t(h)$, where $H_t(h) \equiv \Pr(h_{t+1} \leq h|I_t)$. Then the optimal forecasts for Y and h would respectively satisfy

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c_1(y - f_{t,1}^*) \mathrm{d}F_t(y) = 0 ,$$

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial h_{t,1}} c_2(h - h_{t,1}^*) \mathrm{d}H_t(h) = 0 .$$

It is easy to see that the optimality condition for $f^*_{t,1}$ does not imply the optimality condition for $h^*_{t,1}$ in general. Under some strong conditions on the functional forms of the transformation $h(\cdot)$ and of the two loss functions $c_1(\cdot)$, $c_2(\cdot)$, the above two conditions may coincide. Granger [67] remarks that it would be strange behavior to use the same loss function for Y and $h(Y)$. We leave this for further analysis in a future research.

Density Forecast Evaluation

Most of the classical finance theories, such as asset pricing, portfolio selection and option valuation, aim to model the surrounding uncertainty via a parametric distribution function. For example, extracting information about market participants' expectations from option prices can be considered another form of density forecasting exercise [92]. Moreover, there has also been increasing interest in evaluating forecasting models of inflation, unemployment and output in terms of density forecasts [29]. While evaluating each density forecast model has become versatile since Diebold et al. [35], there has been much less effort in comparing alternative density forecast models.

Given the recent empirical evidence on volatility clustering and asymmetry and fat-tailedness in financial return series, relative adequacy of a given model among alternative models would be useful measure of evaluating forecast models. Deciding on which distribution and/or volatility specification to use for a particular asset is a common task even for finance practitioners. For example, despite the existence of many volatility specifications, a consensus on which model is most appropriate has yet to be reached. As argued in Poon and Granger [125], most of the (volatility) forecasting studies do not produce very conclusive results because only a subset of alternative models are compared, with a potential bias towards the method developed by the authors. Poon and Granger [125] argue that lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task. They wrote (p. 507), "… it seems clear that one form of study that is included is conducted just to support a viewpoint that a particular method is useful. It might not have been submitted for publication if the required result had not been reached. This is one of the obvious weaknesses of a comparison such as this; the papers being prepared for different reasons, use different data sets, many kinds of assets, various intervals between readings, and a variety of evaluation techniques".

Following Diebold et al. [35], it has become common practice to evaluate the adequacy of a forecast model based on the probability integral transform (PIT) of the process with respect to the model's density forecast. If the density forecast model is correctly specified, the PIT follows an i.i.d. uniform distribution on the unit interval and, equivalently, its inverse normal transform follows an i.i.d. normal distribution. We can therefore evaluate a density forecast model by examining the departure of the transformed PIT from this property (i.i.d. and normality). The departure can be quantified by the Kullback-Leibler [97] information criterion, or KLIC, which is the expected logarithmic value of the likelihood ratio (LR) of the transformed PIT and the i.i.d. normal variate. Thus the LR statistic measures the distance of a candidate model to the unknown true model.

Consider a financial return series $\{y_t\}^T_{t=1}$. This observed data on a univariate series is a realization of a stochastic process $\mathbf{Y}^T \equiv \{Y_\tau : \Omega \rightarrow \mathbb{R}, \tau = 1, 2, \dots, T\}$ on a complete probability space $(\Omega, \mathcal{F}_T, P^T_0)$, where $\Omega = \mathbb{R}^T \equiv \times^T_{\tau=1}\mathbb{R}$ and $\mathcal{F}_T = \mathcal{B}(\mathbb{R}^T)$ is the Borel σ-field generated by the open sets of $\mathbb{R}^T$, and the *joint* probability measure $P^T_0(B) \equiv P_0[\mathbf{Y}^T \in B]$, $B \in \mathcal{B}(\mathbb{R}^T)$ completely describes the stochastic process. A sample of size T is denoted as $\mathbf{y}^T \equiv (y_1, \dots, y_T)'$.

Let σ-finite measure ν^T on $\mathcal{B}(\mathbb{R}^T)$ be given. Assume $P^T_0(B)$ is absolutely continuous with respect to ν^T for all $T = 1, 2, \dots$, so that there exists a measurable Radon–Nikodým density $g^T(\mathbf{y}^T) = \mathrm{d}P^T_0/\mathrm{d}\nu^T$, unique up to a set of zero measure-ν^T.

Following White [159], we define a probability model $\mathcal{P}$ as a collection of distinct probability measures on the measurable space $(\Omega, \mathcal{F}_T)$. A probability model $\mathcal{P}$ is said to be correctly specified for $\mathbf{Y}^T$ if $\mathcal{P}$ contains P^T_0. Our goal is to evaluate and compare a set of parametric probability models $\{P^T_\theta\}$, where $P^T_\theta(B) \equiv P_\theta[Y^T \in B]$. Suppose there exists a measurable Radon–Nikodým density $f^T(\mathbf{y}^T) = \mathrm{d}P^T_\theta/\mathrm{d}\nu^T$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is a finite-dimensional vector of parameters and is assumed to be identified on $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^k$. See Theorem 2.6 in White [159].

In the context of forecasting, instead of the joint density $g^T(\mathbf{y}^T)$, we consider forecasting the *conditional* density of $\mathbf{Y}^t$, given the information $\mathcal{F}_{t-1}$ generated by $\mathbf{Y}^{t-1}$. Let $\varphi_t(y_t) \equiv \varphi_t(y_t|\mathcal{F}_{t-1}) \equiv g^t(\mathbf{y}^t)/g^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \dots$ and $\varphi_1(y_1) \equiv \varphi_1(y_1|\mathcal{F}_0) \equiv g^1(\mathbf{y}^1) = g^1(y_1)$. Thus the goal is to forecast the (true, unknown) conditional density $\varphi_t(y_t)$.

For this, we use an one-step-ahead conditional density forecast model $\psi_t(y_t; \boldsymbol{\theta}) \equiv \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) \equiv f^t(\mathbf{y}^t)/f^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \dots$ and $\psi_1(y_1) \equiv \psi_1(y_1|\mathcal{F}_0) \equiv f^1(\mathbf{y}^1) = f^1(y_1)$. If $\psi_t(y_t; \boldsymbol{\theta}_0) = \varphi_t(y_t)$ almost surely for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified, and it is said to be optimal because it dominates all other density forecasts for any loss functions as discussed in the previous section (see [35,67,70]).

In practice, it is rarely the case that we can find an optimal model. As it is very likely that "the true distribution is in fact too complicated to be represented by a simple mathematical function" [133], all the models proposed by different researchers can be possibly misspecified and thereby we regard each model as an approximation to the truth. Our task is then to investigate which density forecast model can approximate the true conditional density most closely. We have to first define a metric to measure the distance of a given model to the truth, and then compare different models in terms of this distance.

The adequacy of a density forecast model can be measured by the conditional Kullback-Leibler [97] Information Criterion (KLIC) divergence measure between two conditional densities,

$$\mathbb{I}_t\,(\varphi : \psi, \boldsymbol{\theta}) = \mathbb{E}_{\varphi_t}[\ln \varphi_t\left(y_t\right) - \ln \psi_t\left(y_t; \boldsymbol{\theta}\right)]\ ,$$

where the expectation is with respect to the true conditional density $\varphi_t\,(\cdot|\mathcal{F}_{t-1})$, $\mathbb{E}_{\varphi_t} \ln \varphi_t\left(y_t|\mathcal{F}_{t-1}\right) < \infty$, and $\mathbb{E}_{\varphi_t} \ln \psi_t\left(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}\right) < \infty$. Following White [159], we define the distance between a density model and the true density as the minimum of the KLIC

$$\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}^*_{t-1}\right) = \mathbb{E}_{\varphi_t}\left[\ln \varphi_t\left(y_t\right) - \ln \psi_t\left(y_t; \boldsymbol{\theta}^*_{t-1}\right)\right]\ ,$$

where $\boldsymbol{\theta}^*_{t-1} = \arg\min \mathbb{I}_t\,(\varphi : \psi, \boldsymbol{\theta})$ is the pseudo-true value of $\boldsymbol{\theta}$ [133]. We assume that $\boldsymbol{\theta}^*_{t-1}$ is an interior point of $\boldsymbol{\Theta}$. The smaller this distance is, the closer the density forecast $\psi_t\left(\cdot|\mathcal{F}_{t-1}; \boldsymbol{\theta}^*_{t-1}\right)$ is to the true density $\varphi_t\,(\cdot|\mathcal{F}_{t-1})$.

However, $\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}^*_{t-1}\right)$ is unknown since $\boldsymbol{\theta}^*_{t-1}$ is not observable. We need to estimate $\boldsymbol{\theta}^*_{t-1}$. If our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we split the data into two parts, one for estimation and the other for out-of-sample validation. At each period t in the out-of-sample period ($t = R+1, \ldots, T$), we estimate the unknown parameter vector $\boldsymbol{\theta}^*_{t-1}$ and denote the estimate as $\hat{\boldsymbol{\theta}}_{t-1}$. Using $\{\hat{\boldsymbol{\theta}}_{t-1}\}^T_{t=R+1}$, we can obtain the out-of-sample estimate of $\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}^*_{t-1}\right)$ by

$$\mathbb{I}_P(\varphi : \psi) \equiv \frac{1}{P} \sum_{t=R+1}^{T} \ln[\varphi_t(y_t)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})]$$

where $P = T - R$ is the size of the out-of-sample period. Note that

$$\mathbb{I}_P(\varphi : \psi) = \frac{1}{P} \sum_{t=R+1}^{T} \ln\left[\varphi_t(y_t)/\psi_t\left(y_t; \boldsymbol{\theta}^*_{t-1}\right)\right] + \frac{1}{P} \sum_{t=R+1}^{T} \ln[\psi_t\left(y_t; \boldsymbol{\theta}^*_{t-1}\right)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})]\ ,$$

where the first term in $\mathbb{I}_P(\varphi : \psi)$ measures model uncertainty (the distance between the optimal density $\varphi_t(y_t)$ and the model $\psi_t\left(y_t; \boldsymbol{\theta}^*_{t-1}\right)$) and the second term measures parameter estimation uncertainty due to the distance between $\boldsymbol{\theta}^*_{t-1}$ and $\hat{\boldsymbol{\theta}}_{t-1}$.

Since the KLIC measure takes on a smaller value when a model is closer to the truth, we can regard it as a loss function and use $\mathbb{I}_P(\varphi : \psi)$ to formulate the loss-differential. The out-of-sample average of the loss-differential between model 1 and model 2 is

$$\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2) = \frac{1}{P} \sum_{t=R+1}^{T} \ln\left[\psi^2_t\left(y_t; \hat{\boldsymbol{\theta}}^2_{t-1}\right)/\psi^1_t\left(y_t; \hat{\boldsymbol{\theta}}^1_{t-1}\right)\right]\ ,$$

which is the ratio of the two predictive log-likelihood functions. With treating model 1 as a benchmark model (for model selection) or as the model under the null hypothesis (for hypothesis testing), $\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2)$ can be considered as a loss function to minimize. To sum up, the KLIC differential can serve as a *loss* function for density forecast evaluation as discussed in Bao, Lee, and Saltoglu [10]. See Corradi and Swanson [31] for the related ideas using different loss functions.

Using the KLIC divergence measure to characterize the extent of misspecification of a forecast model, Bao, Lee, and Saltoglu [10], in an empirical study with the S&P500 and NASDAQ daily return series, find strong evidence for rejecting the Normal-GARCH benchmark model, in favor of the models that can capture skewness in the conditional distribution and asymmetry and long-memory in the conditional variance. Also, Bao and Lee [8] investigate the nonlinear predictability of stock returns when the density forecasts are evaluated/compared instead of the conditional mean point forecasts. The conditional mean models they use for the daily closing S&P500 index returns include the martingale difference model, the linear ARMA models, the STAR and SETAR models, the ANN model, and the polynomial model. Their empirical findings suggest that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models while we find no such evidence for the left tails or the entire distribution.

Conclusions

In this article we have selectively reviewed the state-of-the-art in nonlinear time series models that are useful in forecasting financial variables. Overall financial returns are difficult to forecast, and this may just be a reflection of the efficiency of the markets on processing information. The success of nonlinear time series on producing better forecasts than linear models depends on how persistent the nonlinearities are in the data. We should note that though many of the methodological developments are concerned with the specification of the conditional mean and conditional variance, there is an active area of research investigating other aspects of the conditional density – quantiles, directions, intervals – that seem to be promising from a forecasting point of view.

For a more extensive coverage to complement this review, the readers may find the following additional references useful. Campbell, Lo, and MacKinlay [22], Chapter 12, provides a brief but excellent summary of nonlinear time series models for the conditional mean and conditional variance as well and various methods such as ANN and nonparametric methods. Similarly, the interested readers may also refer to the books and monographs of Granger and Teräsvirta [72], Franses and van Dijk [55], Fan and Yao [52], Tsay [153], Gao [57], and some book chapters such as Stock [139], Tsay [152], Teräsvirta [145], and White [161].

Future Directions

Methodological developments in nonlinear time series have happened without much guidance from economic theory. Nonlinear models are for most part ad hoc specifications that, from a forecasting point of view, are validated according to some statistical loss function. Though we have surveyed some articles that employ some economic rationale to evaluate the model and/or the forecast – bull/bear cycles, utility function, profit/loss function –, there is still a vacuum on understanding why, how, and when nonlinearities may show up in the data.

From a methodological point of view, future developments will focus on multivariate nonlinear time series models and their associated statistical inference. Nonlinear VAR-type models for the conditional mean and high-dimensional multivariate volatility models are still in their infancy. Dynamic specification testing in a multivariate setting is paramount to the construction of a multivariate forecast and though multivariate predictive densities are inherently difficult to evaluate, they are most important in financial economics.

Another area of future research will deal with the econometrics of a data-rich environment. The advent of large databases begs the introduction of new techniques and methodologies that permits the reduction of the many dimensions of a data set to a parsimonious but highly informative set of variables. In this sense, criteria on how to combine information and how to combine models to produce more accurate forecasts are highly desirable.

Finally, there are some incipient developments on defining new stochastic processes where the random variables that form the process are of a symbolic nature, i. e. intervals, boxplots, histograms, etc. Though the mathematics of these processes are rather complex, future developments in this area will bring exciting results for the area of forecasting.

Bibliography

1. Ait-Sahalia Y, Hansen LP (2009) Handbook of Financial Econometrics. Elsevier Science, Amsterdam
2. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39(4):885–905
3. Andersen TG, Bollerslev T, Diebold FX, Labys P (2001) The Distribution of Realized Exchange Rate Volatility. J Amer Stat Assoc 96:42–55
4. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71: 579–625
5. Ang A, Bekaert G (2002) Regime Switccheds in Interest Rates. J Bus Econ Stat 20:163–182
6. Bachelier L (1900) Theory of Speculation. In: Cootner P (ed) The Random Character of Stock Market Prices. MIT Press, Cambridge. (1964) reprint
7. Bai J, Ng S (2007) Forecasting Economic Time Series Using Targeted Predictors. Working Paper, New York University and Columbia University
8. Bao Y, Lee TH (2006) Asymmetric Predictive Abilities of Nonlinear Models for Stock Returns: Evidence from Density Forecast Comparison. Adv Econ 20 B:41–62
9. Bao Y, Lee TH, Saltoglu B (2006) Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check. J Forecast 25(2):101–128
10. Bao Y, Lee TH, Saltoglu B (2007) Comparing Density Forecast Models. J Forecast 26(3):203–225
11. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models. J Royal Stat Soc B 64:853–223
12. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH Models: A Survey. J Appl Econ 21:79–109
13. Bera AK, Higgins ML (1993) ARCH Models: Properties, Estimation, and Testing. J Econ Surv 7:305–366
14. Bollerslev T (1986) Generalized Autoregressive Conditional Heteroskedasticity. J Econ 31:307–327

15. Bollerslev T, Chou RY, Kroner KF (1992) ARCH Models in Finance. J Econ 52:5–59
16. Bollerslev T, Engle RF, Nelson DB (1994) ARCH Models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier Science, Amsterdam
17. Bollerslev T, Engle RF, Wooldridge J (1988) A Capital Asset Pricing Model with Time Varying Covariances. J Political Econ 96:116–131
18. Boero G, Marrocu E (2004) The Performance of SETAR Models: A Regime Conditional Evaluation of Point, Interval, and Density Forecasts. Int J Forecast 20:305–320
19. Breiman L (1996) Bagging Predictors. Machine Learning 24:123–140
20. Breiman L (1996) Heuristics of Instability and Stabilization in Model Selection. Ann Stat 24(6):2350–2383
21. Brooks C, Persand G (2003) Volatility Forecasting for Risk Management. J Forecast 22(1):1–22
22. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, New Jersey
23. Campbell JY, Thompson SB (2007) Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? Harvard Institute of Economic Research, Discussion Paper No. 2084
24. Cai Z, Fan J, Yao Q (2000) Functional-coefficient Regression Models for Nonlinear Time Series. J Amer Stat Assoc 95: 941–956
25. Chen X (2006) Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman JJ, Leamer EE (eds) Handbook of Econometrics, vol 6. Elsevier Science, Amsterdam, Chapter 76
26. Chen R, Tsay RS (1993) Functional-coefficient Autoregressive Models. J Amer Stat Assoc 88:298–308
27. Christofferson PF, Diebold FX (2006) Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics. Manag Sci 52:1273–1287
28. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-linear Models. Int J Forecast 20:169–183
29. Clements MP, Smith J (2000) Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. J Forecast 19:255–276
30. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatter Plots. J Amer Stat Assoc 74: 829–836
31. Corradi V, Swanson NR (2006) Predictive Density Evaluation. In: Granger CWJ, Elliot G, Timmerman A (eds) Handbook of Economic Forecasting. Elsevier, Amsterdam, pp 197–284
32. Corsi F (2004) A Simple Long Memory Model of Realized Volatility. Working Paper, University of Lugano
33. Dahl CM, González-Rivera G (2003) Testing for Neglected Nonlinearity in Regression Models based on the Theory of Random Fields. J Econ 114:141–164
34. Dahl CM, González-Rivera G (2003) Identifying Nonlinear Components by Random Fields in the US GNP Growth. Implications for the Shape of the Business Cycle. Stud Nonlinear Dyn Econ 7(1):art2
35. Diebold FX, Gunther TA, Tay AS (1998) Evaluating Density Forecasts with Applications to Financial Risk Management. Int Econ Rev 39:863–883
36. Diebold FX, Li C (2006) Forecasting the Term Structure of Government Bond Yields. J Econom 130:337–364
37. Diebold FX, Mariano R (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–265
38. Diebold FX, Rudebusch GD (1989) Scoring the Leading Indicators. J Bus 62(3):369–391
39. Ding Z, Granger CWJ, Engle RF (1993) A Long Memory Property of Stock Market Returns and a New Model. J Empir Finance 1:83–106
40. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Returns? J Bank Finance 31:279–296
41. Durland JM, McCurdy TH (1994) Duration-Dependent Transitions in a Markov Model of US GNP Growth. J Bus Econ Stat 12:279–288
42. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least Angle Regression. Ann Stat 32(2):407–499
43. Engel C, Hamilton JD (1990) Long Swings in the Dollar: Are they in the Data and Do Markets Know it? Amer Econ Rev 80(4):689–713
44. Engle RF (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation. Econometrica 50:987–1008
45. Engle RF, Hendry DF, Richard J-F (1983) Exogeneity. Econometrica 51:277–304
46. Engle RF, Kane A, Noh J (1997) Index-Option Pricing with Stochastic Volatility and the Value of Accurate Variance Forecasts. Rev Deriv Res 1:139–157
47. Engle RF, Manganelli S (2004) CaViaR: Conditional autoregressive Value at Risk by regression quantiles. J Bus Econ Stat 22(4):367–381
48. Engle RF, Ng VK, Rothschild M (1990) Asset Pricing with a Factor ARCH Covariance Structure: Empirical Estimates for Treasury Bills. J Econ 45:213–238
49. Engle RF, Russell JR (1998) Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66:1127–1162
50. Fan J, Gijbels I (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London
51. Fan J, Yao Q (1998) Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85: 645–660
52. Fan J, Yao Q (2003) Nonlinear Time Series. Springer, New York
53. Fan J, Yao Q, Cai Z (2003) Adaptive varying-coefficient linear models. J Royal Stat Soc B 65:57–80
54. Fitzenberger B (1997) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. J Econ 82:235–287
55. Franses PH, van Dijk D (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge
56. Gallant AR, Nychka DW (1987) Semi-nonparametric maximum likelihood estimation. Econometrica 55:363–390
57. Gao J (2007) Nonlinear Time Series: Semiparametric and Nonparametric Methods. Chapman and Hall, Boca Raton
58. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get Most out of Returns Data Sampled at Different Frequencies. J Econ 131:59–95
59. Giacomini R, White H (2006) Tests of Conditional Predictive Ability. Econometrica 74:1545–1578
60. Glosten LR, Jaganathan R, Runkle D (1993) On the Relation-

ship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801

61. González-Rivera G (1998) Smooth-Transition GARCH Models. Stud Nonlinear Dyn Econ 3(2):61–78
62. González-Rivera G, Lee TH, Mishra S (2004) Forecasting Volatility: A Reality Check Based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood. Int J Forecast 20(4):629–645
63. González-Rivera G, Lee TH, Mishra S (2008) Jumps in Cross-Sectional Rank and Expected Returns: A Mixture Model. J Appl Econ; forthcoming
64. González-Rivera G, Lee TH, Yoldas E (2007) Optimality of the Riskmetrics VaR Model. Finance Res Lett 4:137–145
65. Gonzalo J, Martíneza O (2006) Large shocks vs. small shocks. (Or does size matter? May be so). J Econ 135:311–347
66. Goyal A, Welch I (2006) A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Working Paper, Emory and Brown, forthcoming in Rev Financ Stud
67. Granger CWJ (1999) Outline of Forecast Theory Using Generalized Cost Functions. Span Econ Rev 1:161–173
68. Granger CWJ (2002) Some Comments on Risk. J Appl Econ 17:447–456
69. Granger CWJ, Newbold P (1986) Forecasting Economic Time Series, 2nd edn. Academic Press, San Diego
70. Granger CWJ, Pesaran MH (2000) A Decision Theoretic Approach to Forecasting Evaluation. In: Chan WS, Li WK, Tong H (eds) Statistics and Finance: An Interface. Imperial College Press, London
71. Granger CWJ, Lee TH (1999) The Effect of Aggregation on Nonlinearity. Econ Rev 18(3):259–269
72. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, New York
73. Guidolin M, Timmermann A (2006) An Econometric Model of Nonlinear Dynamics in the Joint Distribution of Stock and Bond Returns. J Appl Econ 21:1–22
74. Haggan V, Ozaki T (1981) Modeling Nonlinear Vibrations Using an Amplitude-dependent Autoregressive Time Series Model. Biometrika 68:189–196
75. Hamilton JD (1994) Time Series Analysis. Princeton University Press, New Jersey
76. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57:357–384
77. Hamilton JD (1996) Specification Testing in Markov-Switching Time Series Models. J Econ 70:127–157
78. Hamilton JD (2001) A Parametric Approach to Flexible Nonlinear Inference. Econometrica 69:537–573
79. Hamilton JD, Jordà O (2002) A Model of the Federal Funds Target. J Political Econ 110:1135–1167
80. Hansen BE (1996) Inference when a Nuisance Parameter is not Identified under the Null Hypothesis. Econometrica 64: 413–430
81. Hansen PR (2005) A test for superior predictive ability. J Bus Econ Stat 23:365–380
82. Harding D, Pagan A (2002) Dissecting the Cycle: A Methodological Investigation. J Monet Econ 49:365–381
83. Härdle W, Tsybakov A (1997) Local polynomial estimators of the volatility function in nonparametric autoregression. J Econ 81:233–242
84. Harter HL (1977) Nonuniqueness of Least Absolute Values Regression. Commun Stat – Theor Methods A6:829–838
85. Hong Y, Chung J (2003) Are the Directions of Stock Price Changes Predictable? Statistical Theory and Evidence. Working Paper, Department of Economics, Cornell University
86. Hong Y, Lee TH (2003) Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models. Rev Econ Stat 85(4):1048–1062
87. Hong Y, Lee TH (2003b) Diagnostic Checking for Adequacy of Nonlinear Time Series Models. Econ Theor 19(6):1065–1121
88. Hornik K, Stinchcombe M, White H (1989) Multi-Layer Feedforward Networks Are Universal Approximators. Neural Netw 2:359–366
89. Huang H, Tae-Hwy L, Canlin L (2007) Forecasting Output Growth and Inflation: How to Use Information in the Yield Curve. Working Paper, University of California, Riverside, Department of Economics
90. Huang YL, Kuan CM (2007) Re-examining Long-Run PPP under an Innovation Regime Switching Framework. Academia Sinica, Taipei
91. Inoue A, Kilian L (2008) How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US CPI Inflation, forthcoming. J Amer Stat Assoc 103(482):511–522
92. Jackwerth JC, Rubinstein M (1996) Recovering probability distributions from option prices. J Finance 51:1611–1631
93. Judd KL (1998) Numerical Methods in Economics. MIT Press, Cambridge
94. Kanas A (2003) Non-linear Forecasts of Stock Returns. J Forecast 22:299–315
95. Koenker R, Bassett Jr G (1978) Regression Quantiles. Econometrica 46(1):33–50
96. Kuan CM, Huang YL, Tsayn RS (2005) An unobserved component model with switching permanent and transitory innovations. J Bus Econ Stat 23:443–454
97. Kullback L, Leibler RA (1951) On Information and Sufficiency. Ann Math Stat 22:79–86
98. Lee TH, Ullah A (2001) Nonparametric Bootstrap Tests for Neglected Nonlinearity in Time Series Regression Models. J Nonparametric Stat 13:425–451
99. Lee TH, White H, Granger CWJ (1993) Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests. J Econ 56:269–290
100. Lee TH, Yang Y (2006) Bagging Binary and Quantile Predictors for Time Series. J Econ 135:465–497
101. Lettau M, Ludvigson S (2001) Consumption, Aggregate Wealth, and Expected Stock Returns. J Finance 56:815–849
102. Lewellen J (2004) Predicting Returns with Financial Ratios. J Financial Econ 74:209–235
103. Lintner J (1965) Security Prices, Risk and Maximal Gains from Diversification. J Finance 20:587–615
104. Linton O, Whang YJ (2007) A Quantilogram Approach to Evaluating Directional Predictability. J Econom 141:250-282
105. Lo AW, MacKinlay AC (1999) A Non-Random Walk Down Wall Street. Princeton University Press, Princeton
106. Long X, Su L, Ullah A (2007) Estimation and Forecasting of Dynamic Conditional Covariance: A Semiparametric Multivariate Model. Working Paper, Department of Economics, UC Riverside
107. Lopez JA (2001) Evaluating the Predictive Accuracy of Volatility Models. J Forecast 20:87–109
108. Ludvigson S, Ng S (2007) The Empirical Risk Return Relation: A Factor Analysis Approach. J Financ Econ 83:171–222

109. Lundbergh S, Teräsvirta T (2002) Forecasting with smooth transition autoregressive models. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 21
110. Luukkonen R, Saikkonen P, Teräsvirta T (1988) Testing Linearity in Univariate Time Series Models. Scand J Stat 15:161–175
111. Maheu JM, McCurdy TH (2000) Identifying Bull and Bear Markets in Stock Returns. J Bus Econ Stat 18:100–112
112. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3(3):205–228
113. Markowitz H (1959) Portfolio Selection: Efficient Diversification of Investments. John Wiley, New York
114. Marsh IW (2000) High-frequency Markov Switching Models in the Foreign Exchange Market. J Forecast 19:123–134
115. McAleer M, Medeiros MC (2007) A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. J Econ; forthcoming
116. Money AH, Affleck-Graves JF, Hart ML, Barr GDI (1982) The Linear Regression Model and the Choice of p. Commun Stat – Simul Comput 11(1):89–109
117. Nyguist H (1983) The Optimal L_p-norm Estimation in Linear Regression Models. Commun Stat – Theor Methods 12: 2511–2524
118. Nadaraya ÉA (1964) On Estimating Regression. Theor Probab Appl 9:141–142
119. Nelson CR, Siegel AF (1987) Parsimonious Modeling of Yield Curves. J Bus 60:473–489
120. Nelson DB (1991) Conditional Heteroscedasticity in Asset Returns: A New Approach. Econometrica 59(2):347–370
121. Noh J, Engle RF, Kane A (1994) Forecasting Volatility and Option Prices of the S&P 500 Index. J Deriv 17–30
122. Pagan AR, Ullah A (1999) Nonparametric Econometrics. Cambridge University Press, Cambridge
123. Patton AJ, Timmermann A (2007) Testing Forecast Optimality Under Unknown Loss. J Amer Stat Assoc 102(480): 1172–1184
124. Perez-Quiros G, Timmermann A (2001) Business Cycle Asymmetries in Stock Returns: Evidence form Higher Order Moments and Conditional Densities. J Econ 103:259–306
125. Poon S, Granger CWJ (2003) Forecasting volatility in financial markets. J Econ Lit 41:478–539
126. Powell JL (1986) Censored Regression Quantiles. J Econ 32:143–155
127. Priestley MB (1980) State-dependent Models: A General Approach to Nonlinear Time Series Analysis. J Time Ser Anal 1:47–71
128. Raj B, Ullah A (1981) Econometrics: A Varying Coefficients Approach. Croom Helm, London
129. Riskmetrics (1995) Technical Manual, 3rd edn. New York
130. Romano JP, Wolf M (2005) Stepwise multiple testing as formalized data snooping. Econometrica 73:1237–1282
131. Ross S (1976) The Arbitrage Theory of Capital Asset Pricing. J Econ Theor 13:341–360
132. Ruppert D, Wand MP (1994) Multivariate Locally Weighted Least Squares Regression. Ann Stat 22:1346–1370
133. Sawa T (1978) Information Criteria for Discriminating among Alternative Regression Models. Econometrica 46:1273–1291
134. Schwert GW (1990) Stock Volatility and the Crash of '87. Rev Financ Stud 3(1):77–102
135. Sentana E (1995) Quadratic ARCH models. Rev Econ Stud 62(4):639–661
136. Sichel DE (1994) Inventories and the Three Phases of the Business Cycle. J Bus Econ Stat 12:269–277
137. Sharpe W (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. J Finance 19:425–442
138. Stinchcombe M, White H (1998) Consistent Specification Testing with Nuisanse Parameters Present only under the Alternative. Econ Theor 14:295–325
139. Stock JH (2001) Forecasting Economic Time Series. In: Baltagi BP (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, Chapter 27
140. Stock JH, Watson MW (2002) Forecasting Using Principal Components from a Large Number of Predictors. J Amer Stat Assoc 97:1167–1179
141. Stock JH, Watson MW (2006) Forecasting Using Many Predictors. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam
142. Stone CJ (1977) Consistent Nonparametric Regression. Ann Stat 5:595–645
143. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
144. Teräsvirta T (1994) Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models. J Amer Stat Assoc 89:208–218
145. Teräsvirta T (2006) Forecasting economic variables with nonlinear models. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, pp 413–457
146. Teräsvirta T, Anderson H (1992) Characterizing Nonlinearities in Business Cycles using Smooth Transition Autoregressive Models. J Appl Econ 7:119–139
147. Teräsvirta T, Lin CF, Granger CWJ (1993) Power of the Neural Network Linearity Test. J Time Ser Analysis 14:209–220
148. Tong H (1983) Threshold Models in Nonlinear Time Series Analysis. Springer, New York
149. Tong H (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford
150. Trippi R, Turban E (1992) Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance. McGraw-Hill, New York
151. Tsay RS (1998) Testing and Modeling Multivariate Threshold Models. J Amer Stat Assoc 93:1188–1202
152. Tsay RS (2002) Nonlinear Models and Forecasting. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 20
153. Tsay RS (2005) Analysis of Financial Time Series, 2nd edn. Wiley, New York
154. Varian HR (1975) A Bayesian Approach to Real Estate Assessment. In: Fienberg SE, Zellner A (eds) Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage. North Holland, Amsterdam, pp 195–208
155. Watson GS (1964) Smooth Regression Analysis. Sankhya Series A 26:359–372
156. West KD (1996) Asymptotic inference about predictive ability. Econometrica 64:1067–1084
157. West KD, Edison HJ, Cho D (1993) A Utility Based Comparison of Some Models of Exchange Rate Volatility. J Int Econ 35: 23–45

158. White H (1989) An Additional Hidden Unit Tests for Neglected Nonlinearity in Multilayer Feedforward Networks. In: Proceedings of the International Joint Conference on Neural Networks, Washington, DC. IEEE Press, New York, II, pp 451–455
159. White H (1994) Estimation, Inference, and Specification Analysis. Cambridge University Press, Cambridge
160. White H (2000) A Reality Check for Data Snooping. Econometrica 68(5):1097–1126
161. White H (2006) Approximate Nonlinear Forecasting Methods. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, chapter 9
162. Zellner A (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. J Amer Stat Assoc 81:446–451
163. Ziegelmann FA (2002) Nonparametric estimation of volatility functions: the local exponential estimator. Econ Theor 18:985–991

Financial Forecasting, Sensitive Dependence

MOTOTSUGU SHINTANI
Vanderbilt University, Nashville, USA

Article Outline

Glossary

Global Lyapunov exponent A global stability measure of the nonlinear dynamic system. It is a long-run average of the exponential growth rate of infinitesimally small initial deviation and is uniquely determined in the ergodic and stationary case. In this sense, this initial value sensitivity measure does not depend on the initial value. A system with positive Lyapunov exponents is considered chaotic for both deterministic and stochastic cases.

Local Lyapunov exponent A local stability measure based on a short-run average of the exponential growth rate of infinitesimally small initial deviations. Unlike the global Lyapunov exponent, this initial value sensitivity measure depends both on the initial value and the horizon for the average calculation. A smaller local Lyapunov exponent implies a better performance at the point of forecast.

Noise amplification In a stochastic system with the additive noise, the effect of shocks can either grow, remain, or die out with the forecast horizon. If the system is nonlinear, this effect depends both on the initial value and size of the shock. For a chaotic system, the degree of noise amplification is so high that it makes the forecast almost identical to the iid forecast within the next few steps ahead.

Nonlinear impulse response function In a stochastic system with the additive noise, the effect of shocks on the variable in subsequent periods can be summarized in impulse response functions. If the system is linear, the impulse response does not depend on the initial value and its shape is proportional to the size of shocks. If the system is nonlinear, however, the impulse response depends on the initial value, or the history, and its shape is no longer proportional to the size of shocks.

Definition of the Subject

Empirical studies show that there are at least some components in future asset returns that are predictable using information that is currently available. When the linear time series models are employed in prediction, the accuracy of the forecast does not depend on the current return or the initial condition. In contrast, with nonlinear time series models, properties of the forecast error depend on the initial value or the history. The effect of the difference in initial values in a stable nonlinear model, however, usually dies out quickly as the forecast horizon increases. For both deterministic and stochastic cases, the dynamic system is chaos if a small difference in the initial value is amplified at an exponential rate. In a chaotic nonlinear model, the reliability of the forecast can decrease dramatically even for a moderate forecast horizon. Thus, the knowledge of the sensitive dependence on initial conditions in a particular financial time series offers practically useful information on its forecastability. The most frequently used measure of initial value sensitivity is the largest Lyapunov exponent, defined as the long-run average growth rate of the difference between two nearby trajectories. It is a global initial value sensitivity measure in the sense that it contains the information on the global dynamic property of the whole system. The dynamic properties around a single point in the system can be also described using other local measures. Both global and local measures of the sensitive dependence on initial conditions can be estimated nonparametrically from data without specifying the functional form of the nonlinear autoregressive model.

Introduction

When the asset market is efficient, all the information contained in the history of the asset price is already reflected in the current price of the asset. Mathematically, the conditional mean of asset returns becomes independent of the conditioning information set, and thus price changes must be unpredictable (a martingale property). A convenient model to have such a characteristic is a random walk model with independent and identically distributed (iid) increments given by

$$\ln P_t - \ln P_{t-1} = x_t$$

for $t = 0, 1, 2, \ldots$, where P_t is the asset price and x_t is an iid random variable with mean μ_x and variance σ_x^2. When $\mu_x = 0$, the model becomes a random walk without drift, otherwise, it is a random walk with drift μ_x.

Chaos is a nonlinear deterministic process that can generate a random-like fluctuation. In principle, if a purely deterministic model, instead of a random walk process, is used to describe the dynamics of the asset return x_t, all future asset returns should be completely predictable. However, in the case of chaos, a small perturbation can make the performance of a few steps ahead forecast almost identical to that of a random walk forecast. A leading example is the tent map:

$$x_t = 1 - |2x_{t-1} - 1|$$

with some initial value x_0 between 0 and 1. This map almost surely has the uniform distribution $U(0,1)$ as its natural measure, defined as the distribution of a typical trajectory of x_t. This dynamic system thus provides aperiodic trajectory or random-like fluctuation of x_t as the number of iteration increases. By introducing a randomness in the initial value x_0, marginal distribution of x_t approaches the natural measure. This property, referred to as ergodicity, implies that the temporal average of any smooth function of a trajectory x_t, $M^{-1}\sum_{t=0}^{M-1} h(x_t)$, converges to a mathematical expectation $E[h(x_t)] = \int_{-\infty}^{\infty} h(x)\pi(x)\mathrm{d}x$ as M tends to infinity, where the marginal distribution of x_t is expressed in terms of the probability density function (pdf) $\pi(x)$. The natural measure $U(0,1)$ is also a stationary or invariant distribution since the marginal distribution of x_t for any $t \geq 1$, is $U(0,1)$ whenever initial value x_0 follows $U(0,1)$. In this case, the mean μ_x and variance σ_x^2 are 1/2 and 1/12, respectively. Furthermore, almost all the trajectories are second-order white noise in the sense that they have a flat spectrum and zero autocorrelation at all leads and lags.

Therefore, the knowledge of the marginal distribution $\pi(x)$ or spectrum of asset returns cannot be directly used to distinguish between the case of a random walk combined with an iid random variable and the case of the tent map generating the returns. Yet, the two cases have significantly different implications on the predictability of asset returns, at least for the extremely short horizon. When the initial value x_0 is given, using $\mu_x = 1/2$ as a one-period-ahead forecast at $t = 0$ provides a minimum mean square forecast error (MSFE) of $\sigma_x^2 = 1/2$ for the former case. In contrast, using $1 - |2x_0 - 1|$ as the forecast gives zero forecast error for the latter case. With a very tiny perturbation, however, the MSFE of the multiple-period-ahead forecast for the latter case quickly approaches $\sigma_x^2 = 1/2$, which is identical to that of the random walk case.

Another example is the logistic map:

$$x_t = 4x_{t-1}(1 - x_{t-1})$$

with some initial value x_0 between 0 and 1. This map again provides chaotic fluctuation with the natural measure almost surely given by beta distribution $B(1/2, 1/2)$. Provided the same distribution as its stationary or invariant distribution, the mean μ_x and variance σ_x^2 are 1/2 and 1/8, respectively (see [37] for the invariant distribution of the logistic map in general). Again, the random walk model combined with an iid random variable with the same marginal distribution $B(1/2, 1/2)$ is not distinguishable from the logistic map based only on the marginal distribution nor spectra. But the two have very different predictive implications.

The key feature of chaos that is not observed in the iid random variable is the sensitivity of the trajectories to the choice of initial values. This sensitivity can be measured by the Lyapunov exponent which is defined as the average rate of divergence (or convergence) of two nearby trajectories. Indeed, the positivity of the Lyapunov exponent in a bounded dissipative nonlinear system is a widely used formal definition of chaos. To derive this measure for the two examples above, first consider a one-dimensional general nonlinear system

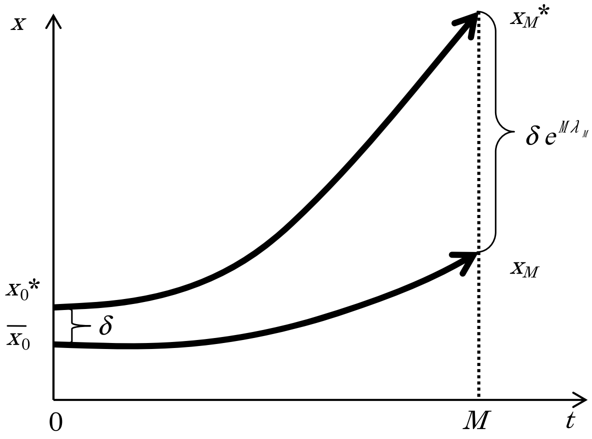
$$x_t = f(x_{t-1})$$

where $f\colon R \to R$ is a continuously differentiable map, with two initial values $x_0 = \overline{x}_0$ and $x_0^* = \overline{x}_0 + \delta$ where δ represents infinitesimal difference in the initial condition. When the distance between two trajectories $\{x_t\}_{t=0}^{\infty}$ and $\{x_t^*\}_{t=0}^{\infty}$ after M steps is measured by the exponential growth rate of δ using $x_M^* - x_M = \delta\exp(M\lambda_M(\overline{x}_0))$, the average of the growth rate in each iteration is given by

$$\lambda_M(\overline{x}_0) = \frac{1}{M}\ln\left|\frac{x_M^* - x_M}{\delta}\right| .$$

Further, let $f^{(M)}$ be the M-fold composition of f. Then from the first order term in the Taylor series expansion of $x_M^* - x_M = f^{(M)}(\overline{x}_0 + \delta) - f^{(M)}(\overline{x}_0)$ around $\overline{x}_0$, combined with the chain rule applied to $df^{(M)}(x)/dx|_{x=\overline{x}_0}$ yields $[f'(x_0)f'(x_1)\cdots f'(x_{M-1})]\delta = [\prod_{t=1}^{M} f'(x_{t-1})]\delta$. Thus, the product $\prod_{t=1}^{M} f'(x_{t-1})$ is the amplifying factor to the initial difference after M periods. Substituting this approximation result into the average growth rate formula yields $\lambda_M(\overline{x}_0) = M^{-1}\sum_{t=1}^{M}\ln|f'(x_{t-1})|$. This measure is called a *local Lyapunov exponent* (of order M) and in general depends on both $\overline{x}_0$ and M (see Fig. 1).

Next, consider the case M tending to infinity. If x_t is ergodic and stationary, $M^{-1}\sum_{t=1}^{M}\ln|f'(x_{t-1})|$ converges



Financial Forecasting, Sensitive Dependence, Figure 1
Lyapunov exponent is an exponential growth rate

to $E[\ln|f'(x_{t-1})|] = \int_{-\infty}^{\infty} \ln|f'(x)|\,\pi(x)\mathrm{d}x$, which does not depend on $\overline{x}_0$. Thus, a *global Lyapunov exponent*, or simply a *Lyapunov exponent*, of a one-dimensional system is defined as $\lambda = \lim_{M\to\infty} \lambda_M(\overline{x}_0)$ or

$$\lambda = \lim_{M\to\infty} \frac{1}{M} \sum_{t=1}^{M} \ln\left|f'(x_{t-1})\right| \ .$$

According to this definition, the computation of the Lyapunov exponent of the tent map is straightforward. Since the tent map is $x_t = 2x_{t-1}$ for $0 \le x_{t-1} \le 1/2$ and $x_t = 2 - 2x_{t-1}$ for $1/2 < x_{t-1} \le 1$, its first derivative $f'(x_{t-1})$ is 2 for $0 \le x_{t-1} \le 1/2$ and -2 for $1/2 < x_{t-1} \le 1$. Using the uniform distribution as its stationary distribution, we have

$$\lambda = \int_0^{1/2} \ln|2|\,\mathrm{d}x + \int_{1/2}^{1} \ln|-2|\,\mathrm{d}x = \ln 2\ (\approx 0.693)\ .$$

Similarly, for the logistic map $x_t = ax_{t-1}(1 - x_{t-1})$ with $a = 4$,

$$\lambda = \int_0^1 \frac{\ln|4 - 8x|}{\pi\sqrt{x(1-x)}}\mathrm{d}x = \ln 4 - \ln 2 = \ln 2\ .$$

Thus, both the tent map and the logistic map with $a = 4$ have a common positive Lyapunov exponent. The value $\ln 2$ implies that, on average, the effect of an initial deviation doubles each time of iteration. Such a rapid rate of divergence is the source of the fact that their trajectories become unpredictable very quickly. Chaos is thus characterized by sensitive dependence on initial conditions measured by a positive Lyapunov exponent.

Let us next consider a simple linear difference equation $x_t = \rho x_{t-1}$ with $|\rho| < 1$. Since its first derivative $f'(x)$ is a constant ρ, not only the Lyapunov exponent but also the local Lyapunov exponent $\lambda_M(\overline{x}_0)$ does not depend on the initial value $\overline{x}_0$. For example, when $\rho = 0.5$, $\lambda = \lambda_M(\overline{x}_0) = -\ln 2$ (≈ -0.693). The logistic map $x_t = ax_{t-1}(1 - x_{t-1})$, can be either chaotic or stable depending on the choice of a. When $a = 1.5$, all the trajectories converge to a point mass at $x_t = 1/3$, where the first derivative is 1/2 thus $\lambda = -\ln 2$. For these two examples, the system has a common negative Lyapunov exponent. In this case, the effect of the initial condition is short-lived and the system is not sensitive to initial conditions. The value $-\ln 2$ implies that, on average, the effect of initial deviation reduces by one half each time of iteration.

Knowing the Lyapunov exponent of the asset returns, or their transformation, thus offers a useful information regarding the predictability of a financial market. In particular, for a system with sensitive dependence (namely, the one with a positive Lyapunov exponent), the performance of a multiple step forecast can worsen quickly as the forecast horizon increases if there are (i) a small uncertainty about the current value at the time of forecast (observation noise) and/or (ii) a small additive noise in the system (system noise).

Lyapunov Exponent and Forecastability

Lyapunov Spectrum

As a global measure of initial value sensitivity in a multidimensional system, the largest Lyapunov exponent and Lyapunov spectrum will first be introduced. For the p-dimensional deterministic nonlinear system,

$$x_t = f\left(x_{t-1}, \ldots, x_{t-p}\right)\ ,$$

where $f\colon R^p \to R$ is continuously differentiable, the (global) largest Lyapunov exponent of the system is defined as

$$\lambda = \lim_{M\to\infty} \frac{1}{2M} \ln\left|\nu_1(\mathsf{T}'_M \mathsf{T}_M)\right|$$

where $\nu_1(\mathsf{T}'_M\mathsf{T}_M)$ is the largest eigenvalue of $\mathsf{T}'_M\mathsf{T}_M$, and $\mathsf{T}_M = J_{M-1} \cdot J_{M-2} \cdot \cdots \cdot J_0$. Here J_{t-1}'s are Jacobian matrices defined as

$$J_{t-1} = \begin{bmatrix} \Delta f_1(\mathsf{X}_{t-1}) & \Delta f_2(\mathsf{X}_{t-1}) & \cdots & \Delta f_{p-1}(\mathsf{X}_{t-1}) & \Delta f_p(\mathsf{X}_{t-1}) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

for $t = 1, \dots, M$, where $\Delta f_j(\mathsf{X}_{t-1}) = \partial f(\mathsf{X}_{t-1})/\partial x_{t-j}$, for $j = 1, \dots, p$, are partial derivatives of the conditional mean function evaluated at $\mathsf{X}_{t-1} = (x_{t-1}, \dots, x_{t-p})'$.

Using an analogy to the one-dimensional case, the local Lyapunov exponent can be defined similarly by $\lambda_M(\mathbf{x}) = (2M)^{-1} \ln \left| \nu_1(\mathsf{T}'_M \mathsf{T}_M) \right|$ with initial value $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \dots, \overline{x}_{-p+1})'$. Note that $(2M)^{-1} \ln \left| \nu_1(\mathsf{T}'_M \mathsf{T}_M) \right|$ reduces to the sum of absolute derivatives in logs used for the one-dimensional case since $(2M)^{-1} \sum_{t=1}^{M} \ln[f'(x_{t-1})^2] = M^{-1} \sum_{t=1}^{M} \ln |f'(x_{t-1})|$.

In the multi-dimensional case, the whole spectrum of Lyapunov exponents can be also considered using ith Lyapunov exponent λ_i, for $i = 1, \dots, p$, defined by replacing the largest eigenvalue ν_1 with the ith largest eigenvalue ν_i. A set of all Lyapunov exponents is called the Lyapunov spectrum. Geometrically, each Lyapunov exponent represents the rate of growth (or contraction) of the corresponding principal axis of a growing (or shrinking) ellipsoid. An attracting set of a dynamic system, or simply the attractor, is defined as the set to which x_t approaches in the limit. The attractor can be a point, a curve, a manifold, or more complicated set. The Lyapunov spectrum contains information on the type of the attractor. For example, a system with all negative Lyapunov exponents has an equilibrium point as an attracting set. To understand this claim, let $\mathbf{x}_{\text{EQ}}$ be an equilibrium point and consider a small initial deviation $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$ from $\mathbf{x}_{\text{EQ}}$. By the linearization of $f : R^p \to R$ at $\mathbf{x}_{\text{EQ}}$, the deviation from $\mathbf{x}_{\text{EQ}}$ after M periods is approximated by $c_1 \eta_1 \exp\{\widetilde{\lambda}_1 M\} + \cdots + c_p \eta_p \exp\{\widetilde{\lambda}_p M\}$, where $\widetilde{\lambda}_i$'s and η_i's are the eigenvalues and eigenvectors of J_{EQ}, respectively, where J_{EQ} is J_{t-1} evaluated at $\mathsf{X}_{t-1} = \mathbf{x}_{\text{EQ}}$, and c_i's are scalar constants. The real part of $\widetilde{\lambda}_i$, denoted by $\text{Re}[\widetilde{\lambda}_i]$, represents the rate of growth (contraction) around the equilibrium point $\mathbf{x}_{\text{EQ}}$ along the direction of η_i if $\text{Re}[\widetilde{\lambda}_i]$ is positive (negative). Thus if $\text{Re}[\widetilde{\lambda}_i] < 0$ for all $i = 1, \dots, p$, $\mathbf{x}_{\text{EQ}}$ is asymptotically stable and is an attractor. Otherwise, $\mathbf{x}_{\text{EQ}}$ is either unstable with $\text{Re}[\widetilde{\lambda}_i] > 0$ for all i, or a saddle point with $\text{Re}[\widetilde{\lambda}_i] > 0$ for some i, provided that none of $\text{Re}[\widetilde{\lambda}_i]$ is zero. In this simple case, ith Lyapunov exponent λ_i corresponds to $\text{Re}[\widetilde{\lambda}_i]$.

Among all the Lyapunov exponents, the largest Lyapunov exponent λ_1, or simply λ, is a key measure to distinguish chaos from other stable systems. By using an analogy to the equilibrium point example, $\lambda > 0$ implies the expansion in the direction of η_1. An attractor requires that the sum of all the Lyapunov exponents be negative since contraction on the whole must be stronger than the expansion. When this condition is met with some positive λ_i's, the system is said to have a strange attractor. Chaos is thus excluded if the largest Lyapunov exponent is not positive. A system with a zero largest Lyapunov exponent implies that the average distance of two orbits (along some directions) is same as their initial deviation, the property often referred to as Lyapunov stability. A zero largest Lyapunov exponent and strictly negative remaining Lyapunov exponents lead to a system with a limit cycle. If only the first two (k) largest Lyapunov exponents are zero, the system has a two-torus (k-torus) attractor. The types of attractors and their relationship to the signs of Lyapunov exponents are summarized in Table 1.

Entropy and Dimension

In a deterministic system with initial value sensitivity, the information on how quickly trajectories separate on the whole has a crucial implication in the predictability. This is because if two different trajectories which are initially indistinguishable become distinguishable after a finite number of steps, the knowledge of the current state is useful in forecasting only up to a finite times ahead. Kolmogorov entropy of the system measures the rate at which information is produced and has a close relationship to the Lyapunov exponents. In general, the sum of all positive Lyapunov exponents provides an upper bound to Kolmogorov entropy, which contains the information on how quickly trajectories separate on the whole. Under some conditions, both the entropy and the sum become identical (see [22,56]). This fact can intuitively be understood as follows. Suppose a system with k positive Lyapunov exponents and an attractor of size L. Here, the size of an attractor roughly refers to the range of an in-

Financial Forecasting, Sensitive Dependence, Table 1
Lyapunov spectrum and attractors

Attractor	Point	Closed curve	k-torus	Strange attractor
Steady state	equilibrium point	limit cycle (periodic)	k-periodic	chaotic
Dimension	0	1	k (integer)	noninteger
Lyapunov exponents	$\lambda_i < 0$ $(i = 1, \dots, p)$	$\lambda_1 = 0$ $\lambda_i < 0$ $(i = 2, \dots, p)$	$\lambda_1 = \cdots = \lambda_k = 0$ $\lambda_i < 0$ $(i = k+1, \dots, p)$	$\lambda_1 > 0$

variant distribution of an attractor, which becomes unpredictable as a result of magnified small initial deviation of size d. Note that the length of the first k principal axes after M steps of iteration is proportional to $\exp(M\sum_{i=1}^{k}\lambda_i)$. From $d\exp(M\sum_{i=1}^{k}\lambda_i) = L$, the expected time M to reach the size of attractor is given by $(1/\sum_{i=1}^{k}\lambda_i)\ln(L/d)$. This result implies that the larger $\sum_{i=1}^{k}\lambda_i$ becomes, the shorter the period during which the path is predictable.

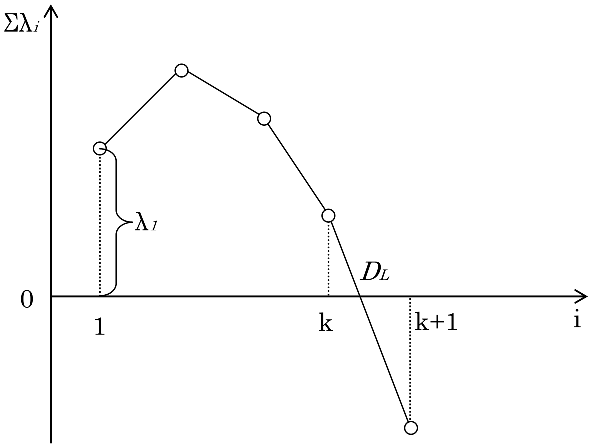
Lyapunov exponents are also closely related to the notion of dimension designed to classify the type of attractors. An equilibrium point has zero dimension. A limit cycle is one-dimensional since it resembles an interval in a neighborhood of any point. A k-torus is k-dimensional since it locally resembles an open subset of R^k. However, the neighborhood of any point of a strange attractor does not resemble any Euclidean space and does not have integer dimension. Among many possibilities of introducing a non-integer dimension, one can consider the Lyapunov dimension, or Kaplan–Yorke dimension, defined as

$$D_L = k + \frac{1}{|\lambda_{k+1}|}\sum_{i=1}^{k}\lambda_i$$

where λ_i is the ith Lyapunov exponent and k is the largest integer for which $\sum_{i=1}^{k}\lambda_i \geq 0$. This definition provides the dimension of zero for an equilibrium point, one for a limit cycle, and k for a k-torus. For a chaotic example, suppose a three-dimensional system with a positive Lyapunov exponent ($\lambda_1 = \lambda_+ > 0$), a zero Lyapunov exponent ($\lambda_2 = 0$), and a negative Lyapunov exponent ($\lambda_3 = \lambda_- < 0$). The Lyapunov dimension D_L is then given by $2+\lambda_+/|\lambda_-|$ which is a fraction that lies strictly between 2 and 3 since an attractor should satisfy $\lambda_+ + \lambda_- < 0$. Likewise, in general, the Lyapunov dimension D_L will be a fraction between k and $k+1$ since $\sum_{i=1}^{k}\lambda_i \leq |\lambda_{k+1}|$ always holds by the definition of k (see Fig. 2).

Since the Lyapunov spectrum contains richer information than the largest Lyapunov exponent alone, several empirical studies reported the Lyapunov spectrum [18], or the transformation such as Kolmogorov entropy and Lyapunov dimension [1,2] of financial time series. However, one must be careful on the interpretation of these quantities since their properties under noisy environment is not rigorously established. In addition, it should be noted that some other forms of the entropy and the dimension can be computed without estimating each Lyapunov exponent separately. For example, [36] recommended using an approximation to Kolmogorov entropy, given by

$$K_2 = \lim_{\substack{\delta\to 0\\ p\to\infty}} \ln\left(\frac{C^p(\delta)}{C^{p+1}(\delta)}\right)$$



Financial Forecasting, Sensitive Dependence, Figure 2
Lyapunov dimension

where $C^p(\delta)$ is the correlation integral defined by

$$C^p(\delta) = \lim_{T\to\infty} \sharp\{(t,s)|\,\|\mathsf{X}_t - \mathsf{X}_s\| < \delta\}/T^2$$

where $\mathsf{X}_t = (x_t, \dots, x_{t-p+1})'$ and $\|\cdot\|$ is a vector norm. The approximation given by K_2 provides a lower bound of Kolmogorov entropy (see [22]). The correlation dimension, a type of dimension, can also be defined as

$$D_C = \lim_{\delta\to 0}\frac{\ln C^p(\delta)}{\ln\delta}\,.$$

Both the K_2 entropy and correlation dimension can be estimated by replacing $C^p(\delta)$ with its sample analogue. In applications to financial time series, these two measures are computed in [30] and [50]. Finally, note that the correlation integral has been used as the basis of the BDS test, a well-established nonlinear dependence test frequently used in economic application, developed by [9]. Formally, the test statistic relies on the sample analogue of $C^p(\delta) - [C^1(\delta)]^p$ and follows normal distribution under the null hypothesis of iid randomness. The BDS test appears to have a good power against the alternative of linear or nonlinear dependence including some low-dimensional chaotic process. Thus, the BDS test is useful in providing the indirect evidence of sensitive dependence and can be complementarily used along with a more direct test based on Lyapunov exponents (see [5] for an example on the comparison between the two approaches).

System Noise and Noisy Chaos

Unlike the data generated from a purely deterministic system, economic and financial data are more likely to be contaminated by noise. There are two main types of random noise used to extend the deterministic model to the stochastic model in the analysis of initial value sensitivity: observation noise and system noise. In the case of the observation noise, or measurement noise, observables are given as the sum of stochastic noise and the unobservables generated from the deterministic model. In contrast, with the system noise, or dynamic noise, observables are generated directly from a nonlinear autoregressive (AR) model. In practice, it is often convenient to introduce the system noise in the additive manner. Theoretically, system noise can make the system to have a unique stationary distribution. Note that for the examples of tent map and logistic map, aperiodic trajectory, or random-like fluctuation, could not be obtained with some choice of initial condition with measure zero. In general, the deterministic system can have infinitely many stationary distributions. However, typically, the presence of additive noise can exclude all degenerate marginal distributions. Furthermore, additive system noise is convenient to generalize the use of the Lyapunov exponents, originally defined in the deterministic system as a measure of sensitive dependence, to the case of a stochastic system.

To see this point, first, consider the following simple linear system with an additive system noise. Adding an iid stochastic error term ε_t, with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$, in the previously introduced linear difference equation leads to a linear AR model of order one,

$$x_t = \rho x_{t-1} + \varepsilon_t \,.$$

The model has a stationary distribution if $|\rho| < 1$. Even if the error term is present, since $f'(x_{t-1}) = \rho$, a one-dimensional Lyapunov exponent can be computed as $\lambda = \ln|\rho| < 0$, the value identical to the case of the deterministic linear difference equation. Thus, the stationarity condition $|\rho| < 1$ in the linear model always implies a negative Lyapunov exponent, while a unit root process $\rho = 1$ implies zero Lyapunov exponent.

Next, consider the introduction of a system noise to a nonlinear system. A general (stationary) nonlinear AR model of order one is defined as

$$x_t = f(x_{t-1}) + \varepsilon_t$$

where $f\colon R \to R$ is a smooth function. For a known unique stationary marginal distribution $\pi(x)$, Lyapunov exponent can be computed as $E[\ln|f'(x_{t-1})|] = \int_{-\infty}^{\infty} \ln|f'(x)|\,\pi(x)\mathrm{d}x$. Thus, by using an analogy of the definition of deterministic chaos, *noisy chaos* can be defined as a stationary nonlinear AR model with a positive Lyapunov exponent. Even if an analytical solution is not available, the value of Lyapunov exponent is typically obtained numerically or by simulation. Similarly, for the multidimensional nonlinear AR model,

$$x_t = f(x_{t-1}, \ldots, x_{t-p}) + \varepsilon_t \,,$$

(noisy) chaos can be defined by a positive largest Lyapunov exponent computed from the Jacobian and the stationary joint distribution of $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})'$. Furthermore, as long as the process has a stationary distribution, for both the chaotic and non-chaotic case, M-period ahead least squares predictor $f_M(\mathbf{x}) \equiv E[x_{t+M}|\mathsf{X}_t = \mathbf{x}]$ and its conditional MSFE $\sigma_M^2(\mathbf{x}) \equiv E[\{x_{t+M} - f_M(\mathbf{x})\}^2|\mathsf{X}_t = \mathbf{x}]$ depend on the initial condition $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ but do not depend on the timing of forecast t.

Noise Amplification

The next issue involves the prediction in the stochastic dynamic system. When additive noise is present in the nonlinear system, the amplification of noise can depend on the initial values and is not necessarily monotonic in horizon. This feature is not unique to the chaotic model but holds for general nonlinear models. However, a small noise is expected to be amplified rapidly in time if the nonlinear system is chaotic.

To understand the process of noise amplification, consider the previously introduced linear AR model of order one with a non-zero coefficient ρ and an initial condition $x_0 = \overline{x}_0$. Then, at the period M,

$$\begin{aligned} x_M &= \rho\{\rho x_{M-2} + \varepsilon_{M-1}\} + \varepsilon_M \\ &= \rho^2 x_{M-2} + \rho\varepsilon_{M-1} + \varepsilon_M \\ &= \rho^M \overline{x}_0 + \varepsilon_M + \cdots + \rho^{M-1}\varepsilon_1 \,. \end{aligned}$$

Since $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M\}$ are not predictable at period 0, the least square M-period ahead predictor is $\rho^M \overline{x}_0$ with its MSFE σ_M^2 given by $\mu_M \sigma^2$ where

$$\mu_M = 1 + \cdots + \rho^{2(M-1)} = 1 + \sum_{j=1}^{M-1} \rho^{2j}$$

is a monotonically increasing proportional factor that does not depend on $\overline{x}_0$. Since $\mu_M > 1$, MSFE is strictly greater than the variance of the noise for all M. However, for a stationary process with $|\rho| < 1$, increments in such a noise amplification become smaller and μ_M converges to $1/(1 - \rho^2)$ as M tends to infinity. Thus, eventually, the MSFE converges to the unconditional variance $\sigma_x^2 = \sigma^2/(1 - \rho^2)$. In a special case with $\rho = 0$, when the asset price have iid increments, the proportional factor becomes 1 for all M giving its MSFE $\sigma_M^2 = \sigma_x^2 = \sigma^2$ for all M.

Suppose, instead, a general nonlinear AR model of order one with an initial condition $x_0 = \overline{x}_0$. In addition, let $|\varepsilon_t| \leq \zeta$ almost surely, where $\zeta > 0$ is a small constant. By Taylor series expansion, for $M \geq 1$,

$$\begin{aligned} x_M &= f\{f(x_{M-2}) + \varepsilon_{M-1}\} + \varepsilon_M \\ &= f^{(2)}(x_{M-2}) + f'\{f(x_{M-2})\}\varepsilon_{M-1} + \varepsilon_M + O(\zeta^2) . \end{aligned}$$

Using the fact that $x_{M-2} = f^{(M-2)}(\overline{x}_0) + O(\zeta)$, and repeating applications of Taylor series expansion,

$$\begin{aligned} x_M &= f^{(2)}(x_{M-2}) + f'\{f^{(M-1)}(\overline{x}_0)\}\varepsilon_{M-1} + \varepsilon_M + O(\zeta^2) \\ &= f^{(M)}(\overline{x}_0) + \varepsilon_M + f'\{f^{(M-1)}(\overline{x}_0)\}\varepsilon_{M-1} + \cdots \\ &\quad + \prod_{k=1}^{M-1} f'\{f^{(k)}(\overline{x}_0)\}\varepsilon_1 + O(\zeta^2) . \end{aligned}$$

Thus the least square M-period ahead predictor is $f^{(M)}(\overline{x}_0)$ with its conditional MSFE given by

$$\sigma_M^2(\overline{x}_0) = \mu_M(\overline{x}_0)\sigma^2 + O(\zeta^3)$$

where

$$\mu_M(\overline{x}_0) = 1 + \sum_{j=1}^{M-1} \left[\prod_{k=j}^{M-1} f'\{f^{(k)}(\overline{x}_0)\} \right]^2 .$$

A comparison of μ_M for the linear model and $\mu_M(\overline{x}_0)$ for the nonlinear model provides some important features of the nonlinear prediction. First, unlike the linear case, the proportional factor now depends not only on the forecast horizon M but also on the initial condition $\overline{x}_0$. Thus, in general, performance of the nonlinear prediction depends on where you are.

Second, $\mu_M(\overline{x}_0)$ does not need to be monotonically increasing with M in nonlinear case. The formula for $\mu_M(\overline{x}_0)$ can be rewritten as

$$\mu_{M+1}(\overline{x}_0) = 1 + \mu_M(\overline{x}_0) f'\{f^{(M)}(\overline{x}_0)\}^2 .$$

Thus, $\mu_{M+1}(\overline{x}_0) < \mu_M(\overline{x}_0)$ is possible when $f'\{f^{(M)}(\overline{x}_0)\}^2 < 1 - 1/\mu_M(\overline{x}_0)$. Therefore, with some initial value and M, the $(M+1)$-period ahead MSFE can be smaller than the M-period ahead MSFE.

Third, and most importantly, unlike the stationary linear model, which imposes the restriction $|\rho| < 1, |f'(x)| > 1$ is possible for a large range of values of x in the nonlinear model even if it has a bounded and stationary distribution. In such a case, $\mu_M(\overline{x}_0)$ can grow rapidly for the moderate or short forecast horizon M. The rapid noise amplification makes the long-horizon forecast very unreliable especially when the model is chaotic. To see this point, it is convenient to rewrite the proportional factor $\mu_M(\overline{x}_0)$ in terms of the local Lyapunov exponent as

$$\mu_M(\overline{x}_0) = 1 + \sum_{j=1}^{M-1} \exp\left\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\right\} .$$

When the local Lyapunov exponent is positive, the proportional factor grows at an exponential rate as M grows. Recall that in the case of iid forecast (random walk forecast in terms of price level), the MSFE σ_M^2 becomes σ_x^2. Likewise, for the chaotic case with infinitesimally small σ^2, the MSFE σ_M^2 reaches σ_x^2 only after a few steps even if the MSFE is close to zero for the one-step ahead forecast. Thus, the global Lyapunov exponent or other local measures of sensitive dependence contain important information on the predictability in the nonlinear time series framework.

Nonparametric Estimation of the Global Lyapunov Exponent

Local Linear Regression

The measures of initial value sensitivity can be computed from the observed data. Since the Lyapunov exponent is by definition the average growth rate of initial deviations between two trajectories, it can be directly computed by finding pairs of neighbors and then averaging growth rates of the subsequent deviations of such pairs [77]. This 'direct' method, however, provides a biased estimator when there is a random component in the system [51]. A modified regression method proposed by [63] is considered more robust to the presence of measurement noise but not necessarily when the system noise is present. A natural approach to compute the Lyapunov exponent in the nonlinear AR model framework is to rely on the estimation of the nonlinear conditional mean function $f: R^p \rightarrow R$. For example, based on an argument similar to the deterministic case, the noisy logistic map, $x_t = ax_{t-1}(1 - x_{t-1}) + \varepsilon_t$, can be either chaotic or stable depending on the value of the parameter a. The first derivative $f'(x) = a - 2ax$ can be evaluated at each data point once an estimate of a is provided. Thus, the parametric approach in the estimation of Lyapunov exponents has been considered in some cases (e. g., [7]). In practice, however, information on the functional form is rarely available and the nonparametric approach is a reasonable alternative. In principle, any nonparametric estimator can be used to estimate the function f and its partial derivatives in the nonlinear AR model,

$$x_t = f\left(x_{t-1}, \ldots, x_{t-p}\right) + \varepsilon_t$$

where f is smooth and ε_t is a martingale difference sequence with $E[\varepsilon_t | x_{t-1}, x_{t-2}, \dots] = 0$ and $E[\varepsilon_t^2 | x_{t-1}, x_{t-2}, \dots] = \sigma^2(x_{t-1}, \dots, x_{t-p}) = \sigma^2(\mathbf{x})$. To simplify the discussion, here, the one based on a particular type of the kernel regression estimator is explained in detail. Methods based on other types of nonparametric estimators will be later mentioned briefly (see, for example, [27], on the nonparametric approach in time series analysis).

The local linear estimator of the conditional mean function and its first partial derivatives at a point $\mathbf{x}$ can be obtained by minimizing the weighted least squares criterion $\sum_{t=1}^{T}(x_t - \beta_0 - \beta_1'(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$, where H is the $d \times d$ bandwidth matrix, K is d-variate kernel function such that $\int K(u)du = 1$, and $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$. For example, the standard p-variate normal density

$$K(u) = \frac{1}{2\pi^{-p/2}} \exp(-||u||^2/2)$$

with H given by hI_p where h is a scalar bandwidth and I_p is an identity matrix of order p, can be used in the estimation. The solution to the minimization problem is given by $\widehat{\beta}(x) = (\mathsf{X}_x' \mathbf{W}_x \mathsf{X}_x)^{-1} \mathsf{X}_x' \mathbf{W}_x \mathbf{Y}$ where

$$\mathsf{X}_x = \begin{bmatrix} 1 & (\mathsf{X}_0 - \mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathsf{X}_{T-1} - \mathbf{x})' \end{bmatrix},$$

$\mathbf{Y} = (x_1, \dots, x_T)'$ and $W_x = \operatorname{diag}\{K_H(\mathsf{X}_0 - \mathbf{x}), \dots, K_H(\mathsf{X}_{T-1} - \mathbf{x})\}$. The local linear estimator of the nonlinear function $f(\mathbf{x})$ and its first derivatives $(\partial f)/(\partial x_{t-j})(\mathbf{x})$ for $j = 1, \dots, p$ are given by $\widehat{\beta}_0(\mathbf{x}) = \widehat{f}(\mathbf{x})$ and

$$\widehat{\beta}_1(\mathbf{x}) = \begin{bmatrix} \widehat{\beta}_{11}(\mathbf{x}) \\ \vdots \\ \widehat{\beta}_{1p}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \Delta\widehat{f}_1(\mathbf{x}) \\ \vdots \\ \Delta\widehat{f}_p(\mathbf{x}) \end{bmatrix},$$

respectively. [22] and [21] proposed a method, known as the 'Jacobian' method, to estimate the Lyapunov exponent by substituting $\Delta f_i(\mathbf{x})$ in the Jacobian formula by its nonparametric estimator $\Delta\widehat{f}_i(\mathbf{x})$. It should be noted that, in general, the "sample size" T used for estimating Jacobian $\widehat{J}_t$ and the "block length" M, which is the number of evaluation points used for estimating the Lyapunov exponent, can be different. Formally, the Lyapunov exponent estimator of λ is given by

$$\widehat{\lambda}_M = \frac{1}{2M} \ln \nu_1 \left(\widehat{\mathsf{T}}_M' \widehat{\mathsf{T}}_M \right),$$

$$\widehat{\mathsf{T}}_M = \prod_{t=1}^{M} \widehat{J}_{M-t} = \widehat{J}_{M-1} \cdot \widehat{J}_{M-2} \cdot \dots \cdot \widehat{J}_0 ,$$

where

$$\widehat{J}_{t-1} = \begin{bmatrix} \Delta\widehat{f}_1(\mathsf{X}_{t-1}) & \Delta\widehat{f}_2(\mathsf{X}_{t-1}) & \cdots & \Delta\widehat{f}_{p-1}(\mathsf{X}_{t-1}) & \Delta\widehat{f}_p(\mathsf{X}_{t-1}) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

for $t = 0, 1, \dots, M-1$, where $\Delta\widehat{f}_j(\mathbf{x})$ is a nonparametric estimator of $\Delta f_j(\mathbf{x}) = \partial f(\mathbf{x}) / \partial x_{t-j}$ for $j = 1, \dots, p$.

As an estimator for the global Lyapunov exponent, setting $M = T$ gives the maximum number of Jacobians and thus the most accurate estimation can be expected. Theoretically, however, it is often convenient to have a block length M smaller than T. For a fixed M, with T tends to infinity, $\widehat{\lambda}_M$ is a consistent estimator of the local Lyapunov exponent with initial value $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \dots, \overline{x}_{-p+1})'$ (see [48]). In case both M and T increase with M/T tends to zero, $\widehat{\lambda}_M$ is still a consistent estimator of the global Lyapunov exponent.

Statistical Inference on the Sign of Lyapunov Exponent

Since the positive Lyapunov exponent is the condition that distinguishes the chaotic process from the stable system without high initial value sensitivity, conducting the inference regarding the sign of the Lyapunov exponent is often of practical interest. For such inference, a consistent standard error formula for $\widehat{\lambda}_M$ is available. Under the condition that M grows at a sufficiently slow rate, a standard error can be computed by $\sqrt{\widehat{\Phi}/M}$ where

$$\widehat{\Phi} = \sum_{j=-M+1}^{M-1} w(j/S_M)\widehat{\gamma}(j)$$

$$\text{with} \quad \widehat{\gamma}(j) = \frac{1}{M} \sum_{t=|j|+1}^{M} \widehat{\eta}_t \widehat{\eta}_{t-|j|} ,$$

$$\widehat{\eta}_t = \widehat{\xi}_t - \widehat{\lambda}_M \quad \text{with} \quad \widehat{\xi}_t = \frac{1}{2} \ln \left(\frac{\nu_1 \left(\widehat{\mathsf{T}}_t' \widehat{\mathsf{T}}_t \right)}{\nu_1 \left(\widehat{\mathsf{T}}_{t-1}' \widehat{\mathsf{T}}_{t-1} \right)} \right)$$

$$\text{for} \quad t \geq 2 \quad \text{and} \quad \widehat{\xi}_1 = \frac{1}{2} \ln \nu_1 \left(\widehat{\mathsf{T}}_1' \widehat{\mathsf{T}}_1 \right) ,$$

where $w(u)$ and S_M denote a kernel function and a lag truncation parameter, respectively (see [67,68,74]). An example of $w(u)$ is the triangular (Bartlett) kernel given by $w(u) = 1 - |u|$ for $|u| < 1$ and $w(u) = 0$, otherwise. The lag truncation parameter S_M should grow at a rate slower than the rate of M.

The procedure above relies on the asymptotic normality of the Lyapunov exponent estimator. Therefore, if the number of Jacobians, M, is not large, an approximation by the normal distribution may not be appropriate. An alternative approach to computing the standard error is to use the resample methods, such as bootstrapping or subsampling. See [32,35] and [79] for the applications of resampling methods to the evaluation of the global Lyapunov exponent estimates.

Consistent Lag Selection

Performance of the nonparametric Lyapunov exponent estimator is often influenced by the choice of lag length p in the nonlinear AR model when the true lag is not known in practice. To see this point, artificial data is generated from a noisy logistic map with an additive system error given by

$$x_t = ax_{t-1}(1 - x_{t-1}) + \sigma(x_{t-1})\varepsilon_t$$

where $\varepsilon_t \sim$ iid $U(-1/2, 1/2)$ and $\sigma(x_{t-1}) = 0.5 \times \min\{ax_{t-1}(1 - x_{t-1}), 1 - ax_{t-1}(1 - x_{t-1})\}$. Note that the conditional heteroskedasticity function $\sigma(x)$ here ensures that the process x_t is restricted to the unit interval $[0, 1]$. When $a = 4.0$, the system has a positive Lyapunov exponent 0.699. Figure 3 shows an example of a sample path from a deterministic logistic map (left) and a noisy logistic map with the current specification of an error term (right). When $a = 1.5$, the system has a negative Lyapunov exponent -0.699. Table 2 reports the mean and median of nonparametric estimates of Lyapunov exponents using the lags from 1 to 4, $M = T = 50$, based on 1,000 replications.

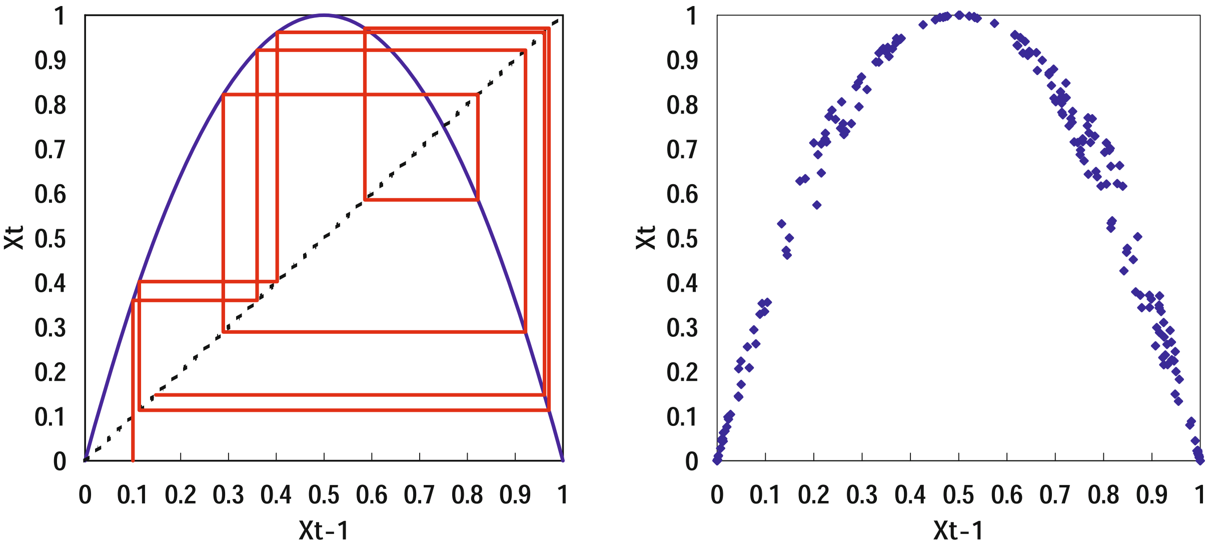
Financial Forecasting, Sensitive Dependence, Table 2
Lyapunov exponent estimates when $T = 50$: logistic map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Logistic map with $a = 4.0$ (true $\lambda = 0.699$)	Mean	0.694	0.706	0.713	0.720
	Median	0.696	0.704	0.710	0.715
Logistic map with $a = 1.5$ (true $\lambda = -0.699$)	Mean	-0.560	-0.046	0.115	0.179
	Median	-0.661	-0.152	0.060	0.149

The simulation results show that overfitting has relatively small effect when the true Lyapunov exponent is positive. On the other hand, in case of negative Lyapunov exponent, the upward bias caused by including redundant lags in the nonparametric regression can result in positive Lyapunov exponent estimates. Therefore, when the true lag length of the system is not known, lag selection procedure will be an important part of the analysis of sensitive dependence.

There are several alternative criteria that are designed to select lag length p in the nonparametric kernel autoregressions. With respect to lag selection in the nonparametric analysis of chaos, [15] suggested minimizing the cross-validation (CV) defined by

$$\widehat{CV}(p) = T^{-1}\sum_{t=1}^{T}\left\{x_t - \widehat{f}_{-(t-1)}(\mathrm{X}_{t-1})\right\}^2 W^2(\mathrm{X}_{t-1})$$



Financial Forecasting, Sensitive Dependence, Figure 3
Logistic map and noisy logistic map

Financial Forecasting, Sensitive Dependence, Table 3
Frequencies of selected lags when $T = 50$: logistic map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Logistic map with $a = 4.0$ (true $\lambda = 0.699$)	CV	0.989	0.011	0.000	0.000
	FPE	0.998	0.002	0.000	0.000
	CFPE	1.000	0.000	0.000	0.000
Logistic map with $a = 1.5$ (true $\lambda = -0.699$)	CV	0.697	0.168	0.080	0.055
	FPE	0.890	0.085	0.017	0.008
	CFPE	0.989	0.011	0.000	0.000

where $\widehat{f}_{-(t-1)}(\mathsf{X}_{t-1})$ is the leave-one-out estimator evaluated at X_{t-1} and $W^2(\mathbf{x})$ is a weight function. [70] suggested minimizing the nonparametric version of the final prediction error (FPE) defined by

$$\widehat{\text{FPE}}(p) = T^{-1}\sum_{t=1}^{T}\left\{x_t - \widehat{f}(\mathsf{X}_{t-1})\right\}^2 W^2(\mathsf{X}_{t-1})$$
$$+ \frac{2}{Th^p}K(0)^p T^{-1}\sum_{t=1}^{T}\{x_t - \widehat{f}(\mathsf{X}_{t-1})\}^2$$
$$\cdot\, W^2(\mathsf{X}_{t-1})/\widehat{\pi}(\mathsf{X}_{t-1})$$

where $\widehat{\pi}(\mathbf{x})$ is a nonparametric joint density estimator at $\mathbf{x}$. [71] proposed a modification to the FPE to prevent overfitting in a finite sample with a multiplicative correction term $\{1+p(T-p+1)\}^{-4/(p+4)}$. All three nonparametric criteria, the CV, FPE, and the corrected version of the FPE (CFPE) are proved to be consistent lag selection criteria so that the probability of selecting the correct p converges to one as T increases. Table 3 reports frequencies of selected lags based on these criteria among 1,000 iterations.

The simulation results show that all the lag selection criteria perform reasonably well when the data is generated from a noisy logistic map.

While a noisy logistic map has the nonlinear AR(1) form, it should be informative to examine the performance of the procedures when the true process is the AR model of a higher lag order. [15] considered a nonlinear AR(2) model of the form,

$$x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2} + \varepsilon_t$$

where $\varepsilon_t \sim$ iid $U(-0.01, 0.01)$. This is a noisy Hénon map with a positive Lyapunov exponent, $\lambda = 0.409$. Table 4 shows the mean and median of 1,000 nonparametric estimates of Lyapunov exponents using the lags from 1 to 4, $M = T = 50$, when the data is artificially generated from this higher order noisy chaos process.

As in the finding from a chaotic logistic map example, estimates do not seem to be very sensitive to the choice of lags. The results on the lag selection criteria are provided in Table 5.

Financial Forecasting, Sensitive Dependence, Table 4
Lyapunov exponent estimates when $T = 50$: Hénon map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Hénon map (true $\lambda = 0.409$)	Mean	0.411	0.419	0.424	0.431
	Median	0.407	0.423	0.427	0.425

Financial Forecasting, Sensitive Dependence, Table 5
Frequencies of selected lags when $T = 50$: Hénon map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Hénon map (true $\lambda = 0.409$)	CV	0.006	0.740	0.250	0.004
	FPE	0.028	0.717	0.253	0.002
	CFPE	0.043	0.762	0.194	0.001

The table shows that frequencies of selecting the true lag ($p = 2$) becomes less than in the case of the chaotic logistic map in Table 3. However, the performance of CV improves when it is compared to the case of stable logistic map.

The results from this small-scale simulation exercise show that when the true lag length is not known, combining the automatic lag selection method with Lyapunov exponent estimation is recommended in practice.

Other Nonparametric Estimators

In addition to the class of kernel regression estimators, which includes Nadaraya–Watson, local linear or local polynomial estimators, other estimators have also been employed in the estimation of the Lyapunov exponent. With the kernel regression method, Jacobians are evaluated using a local approximation to the nonlinear function at the lagged point X_{t-1}. Another example of the local smoothing method used in Lyapunov exponent estimation is the local thin-plate splines suggested by [51,54]. The local estimation method, however, is subject to the data sparseness problem in the high-dimensional system. Alternatively, Jacobians can be evaluated using a global approximation to the unknown function. As a global estimation method, a global spline function may be used to smooth all the available sample. However, the most frequently used global method in Lyapunov exponent estimation in practice is the neural nets ([2,18,68], among others). A single hidden-layer, feedforward neural network is given by

$$f(\mathsf{X}_{t-1}) = \beta_0 + \sum_{j=1}^{k}\beta_j\psi(a_j'\mathsf{X}_{t-1} + b_j)$$

where ψ is an activation function (most commonly a logistic distribution function) and k is a number of hidden units. The neural network estimator $\widehat{f}$ can be obtained by minimizing the (nonlinear) least square criterion. Jacobians are then evaluated using the analytical first derivative of neural net function. Compared to other functional approximations, the neural net form is less sensitive to increasing lag length, p. Thus, it has a merit in terms of the effective sample size.

Four Local Measures of Sensitive Dependence

Local Lyapunov Exponent

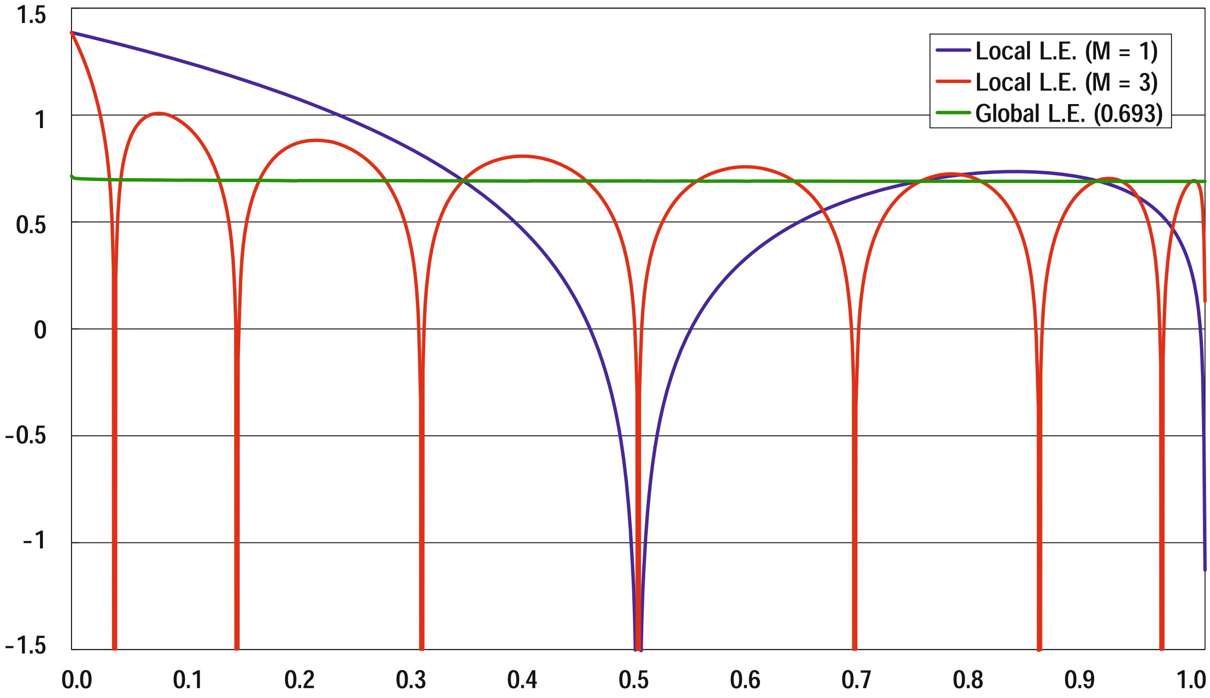
The global Lyapunov exponent measures the initial value sensitivity of long horizon forecast. For the ergodic and stationary case, this initial value sensitivity measure does not depend on the initial value. By definition, the global Lyapunov exponent is the limit of the local Lyapunov exponent when its order M tends to infinity. Unlike the global Lyapunov exponent, the local Lyapunov exponent is a function of an initial value and thus the initial value sensitivity of the short-term forecast depends on where you are. In this sense, local measures of sensitive dependence contain more detailed information on the predictability in the nonlinear dynamic system.

Recall that both the deterministic tent map and the logistic map with $a = 4.0$ have a common positive Lyapunov exponent 0.693. Thus in terms of long-horizon predictability, two processes have exactly the same degree of initial value sensitivity. Yet, in terms of short term forecast, it is possible that predictability at the same point differs among two processes.

The sign of local Lyapunov exponents of the single process can also be different in some range of initial values. Figure 4 shows the local Lyapunov exponents of the deterministic logistic map with $a = 4.0$ for different values of M. Consistent with the definition, as M grows, it approaches to a flat line at the value of 0.693. However, when M is finite, there is a range of initial values associated with a negative local Lyapunov exponent. Within such a range of initial values, sensitive dependence is low and predictability is high even if it is a globally chaotic process.

Analysis of local Lyapunov exponent is also valid in the presence of noise. Studies by [4,48,78], among others, investigate the properties of the local Lyapunov exponent in a noisy system.

The local Lyapunov exponent can be estimated nonparametrically from data using the following procedure. First, obtain the nonparametric Jacobian estimate $\widehat{J}_{t-1}$ for



Financial Forecasting, Sensitive Dependence, Figure 4
Local and global Lyapunov exponents of logistic map

each t using a full sample, as in the case of global Lyapunov exponent estimation. Second, choose a single horizon M of interest. Third, choose the p-dimensional initial value $\mathbf{x} = (x_{t^*}, x_{t^*-1}, \ldots, x_{t^*-p+1})'$ from the data subsequence $\{x_t\}_{t=-p+1}^{T-M}$. Finally, the local Lyapunov exponent estimator at $\mathbf{x}$ is given by $\widehat{\lambda}_M(\mathbf{x}) = (2M)^{-1} \ln \nu_1(\widehat{\mathsf{T}}'_M \widehat{\mathsf{T}}_M)$ where $\widehat{\mathsf{T}}_M = \prod_{t=t^*}^{t^*+M} \widehat{J}_{M-t}$.

While the local Lyapunov exponent is a simple and straightforward local measure of the sensitive dependence, three other useful local measures will be introduced below.

Nonlinear Impulse Response Function

The impulse response function (IRF) is a widely used measure of the persistence effect of shocks in the analysis of economic time series. Here, it is useful to view the IRF as the difference between the two expected future paths: one with and the other without a shock occurred at the current period. When the shock, or the initial deviation, is very small, the notion of impulse responses is thus closely related to the concept of sensitive dependence on initial conditions. To verify this claim, a simple example of a one-dimensional linear IRF is first provided below, followed by the generalization of the IRF to the case of nonlinear time-series model.

For a linear AR model of order one, $x_t = \rho x_{t-1} + \varepsilon_t$, the M-period ahead IRF to a unit shock is defined as

$$\mathrm{IRF}_M = \rho^M .$$

Let $\{x_t^*\}_{t=0}^{\infty}$ be a sample path that contains a single unit shock whereas $\{x_t\}_{t=0}^{\infty}$ is a sample path without any shock. Also let $x_0 = \overline{x}_0$ be an initial condition for the latter path. Then, this linear IRF can be interpreted in two ways. One interpretation is the sequence of the responses to a shock defined to increase one unit of x_0 at time 0 ($x_1 = \rho\overline{x}_0, x_1^* = \rho(\overline{x}_0 + 1), x_1^* - x_1 = \rho, \ldots, x_M^* - x_M = \rho^M$). In this case, the initial value of x_t^* is given as $x_0^* = \overline{x}_0 + 1$, so the shock can be simply viewed as the deviation of two paths at the initial condition. The other interpretation is the sequence of the responses to a shock defined to increase one unit of x_1 at time 1 ($x_1 = \rho\overline{x}_0, x_1^* = \rho\overline{x}_0 + 1, x_1^* - x_1 = 1, \ldots, x_{M+1}^* - x_{M+1} = \rho^M$). In contrast to the first case, two paths have a common initial condition $x_0^* = x_0 = \overline{x}_0$, but the second path is perturbed as if a shock of $\varepsilon_1 = 1$ is realized at time 1 through the dynamic system of $x_t = \rho x_{t-1} + \varepsilon_t$. In either interpretation, however, IRF_M is the difference between x_t^* and x_t at exactly M-period after the shock has occurred and the IRF does not depend on the initial condition $\overline{x}_0$. In addition, the shape of IRF is preserved even if we replace the unit shock with a shock of size δ. The IRF becomes $\rho^M\delta$ and thus the IRF to a unit shock can be considered as a ratio of $\rho^M\delta$ to δ or the normalized IRF.

In the linear framework, the choice between the two interpretations does not matter in practice since the two cases yield exactly the same IRF. However, for nonlinear models, two alternative interpretations lead to different definitions of the IRF. Depending on the objective of the analysis, one may use the former version [31] or the latter version [42,58] of the nonlinear IRFs. The M-period ahead nonlinear impulse response based on the first interpretation considered by [31] is defined as

$$\begin{aligned}\mathrm{IRF}_M(\delta, \mathbf{x}) &= E\left[x_{t+M-1} | \mathsf{X}_{t-1} = \mathbf{x}^*\right] \\ &\quad - E\left[x_{t+M-1} | \mathsf{X}_{t-1} = \mathbf{x}\right] \\ &= E\left[x_M | \mathsf{X}_0 = \mathbf{x}^*\right] - E\left[x_M | \mathsf{X}_0 = \mathbf{x}\right] \\ &= f_M(\mathbf{x}^*) - f_M(\mathbf{x})\end{aligned}$$

where $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})', \mathbf{x}^* = (\overline{x}_0 + \delta, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$. Unlike the linear IRF, the nonlinear IRF depends on the size of shock δ and the initial condition (or the history) $\mathsf{X}_0 = \mathbf{x}$. Interestingly, the partial derivative $\Delta f_{M,1}(\mathbf{x}) = \partial f_M(\mathbf{x}) / \partial x_{t-1}$ corresponds to normalized IRF (proportional to the nonlinear IRF) for small δ since

$$\begin{aligned}&\lim_{\delta \to 0} \frac{\mathrm{IRF}_M(\delta, \mathbf{x})}{\delta} \\ &= \lim_{\delta \to 0} \frac{f_M(\overline{x}_0 + \delta, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})}{\delta} \\ &\qquad - \frac{f(\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})}{\delta} = \Delta f_{M,1}(\mathbf{x}) .\end{aligned}$$

In the one-dimensional case, the IRF simplifies to

$$\begin{aligned}\mathrm{IRF}_M(\delta, \overline{x}_0) &= E\left[x_{t+M-1} | x_{t-1} = \overline{x}_0 + \delta\right] \\ &\quad - E\left[x_{t+M-1} | x_{t-1} = \overline{x}_0\right] \\ &= E\left[x_M | x_0 = \overline{x}_0 + \delta\right] - E\left[x_M | x_0 = \overline{x}_0\right] \\ &= f_M(\overline{x}_0 + \delta) - f_M(\overline{x}_0) .\end{aligned}$$

The first derivative $f'_M(x)$, thus corresponds to the IRF to an infinitesimally small deviation since

$$\lim_{\delta \to 0} \frac{\mathrm{IRF}_M(\delta, \overline{x}_0)}{\delta} = \lim_{\delta \to 0} \frac{f_M(\overline{x}_0 + \delta) - f(\overline{x}_0)}{\delta} = f'_M(\overline{x}_0).$$

Recall that $\lambda_M(\overline{x}_0) = M^{-1} \ln |\prod_{t=1}^{M} f'(x_{t-1})|$. If $\prod_{t=1}^{M} f'(x_{t-1})$ can be approximated by $f'_M(\overline{x}_0)$, both normalized IRF and the local Lyapunov exponent contain the same information regarding the initial value sensitivity.

Next, based on the second interpretation, IRF can be alternatively defined as

$$\begin{aligned} \mathrm{IRF}^*_M(\delta,\mathbf{x}) &= E\left[x_{t+M-1}|x_t = f(\mathbf{x})+\delta, \mathsf{X}_{t-1}=\mathbf{x}\right] \\ &\quad - E(x_{t+M-1}|\mathsf{X}_{t-1}=\mathbf{x}) \\ &= E\left[x_M|x_1 = f(\mathbf{x})+\delta, \mathsf{X}_0=\mathbf{x}\right] \\ &\quad - E(x_M|\mathsf{X}_0=\mathbf{x}) \\ &= f_{M-1}(\mathbf{x}^*) - f_M(\mathbf{x}) \end{aligned}$$

where $\mathsf{X}_{t-1} = (x_{t-1},\ldots,x_{t-p})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$ and $\mathbf{x}^* = (f(\mathbf{x})+\delta, \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2})'$. This version of nonlinear IRF is sometimes referred to as the generalized impulse response function [42,58]. Using the fact that

$$f_M(\mathbf{x}) = f_{M-1}\left(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2}\right) ,$$

the equivalence of the partial derivative $\Delta f_{M-1,1}(\mathbf{x}) = \partial f_{M-1}(\mathbf{x})/\partial x_{t-1}$ and the small deviation IRF can be also shown as

$$\begin{aligned} &\lim_{\delta\to 0}\frac{\mathrm{IRF}^*_M(\delta,\mathbf{x})}{\delta} \\ &\quad= \lim_{\delta\to 0}\frac{f_{M-1}\left(f(\mathbf{x})+\delta, \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2}\right)}{\delta} \\ &\qquad - \frac{f_{M-1}\left(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2}\right)}{\delta} \\ &\quad= \Delta f_{M-1,1}(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2}) \,. \end{aligned}$$

In the one dimensional case, the IRF formula reduces to

$$\begin{aligned} \mathrm{IRF}^*_M(\delta,\overline{x}_0) &= E\left[x_{t+M-1}|x_t = f(\overline{x}_0)+\delta, x_{t-1}=\overline{x}_0\right] \\ &\quad - E(x_{t+M-1}|x_{t-1}=\overline{x}_0) \\ &= E\left[x_M|x_1 = f(\overline{x}_0)+\delta, x_0=\overline{x}_0\right] \\ &\quad - E(x_M|x_0=\overline{x}_0) \\ &= E\left[x_{M-1}|x_0 = f(\overline{x}_0)+\delta\right] \\ &\quad - E(x_M|x_0=\overline{x}_0) \\ &= f_{M-1}(f(\overline{x}_0)+\delta) - f_M(\overline{x}_0) \\ &= f_{M-1}(f(\overline{x}_0)+\delta) - f_{M-1}(f(\overline{x}_0)) \,. \end{aligned}$$

Similarly, the small deviation IRF is given by

$$\begin{aligned} &\lim_{\delta\to 0}\frac{\mathrm{IRF}^*_M(\delta,\overline{x}_0)}{\delta} \\ &\quad= \lim_{\delta\to 0}\frac{f_{M-1}(f(\overline{x}_0)+\delta) - f_{M-1}(f(\overline{x}_0))}{\delta} \\ &\quad= f'_{M-1}(f(\overline{x}_0)) \,. \end{aligned}$$

The nonlinear impulse response function can be estimated nonparametrically without specifying the functional form by an analogy to Lyapunov exponent estimation (see [72] and [66]). Instead of minimizing $\sum_{t=1}^{T}(x_t - \beta_0 - \beta_1'(\mathsf{X}_{t-1}-\mathbf{x}))^2 K_H(\mathsf{X}_{t-1}-\mathbf{x})$, the local linear estimator of M-period ahead predictor $f_M(\mathbf{x})$ and its partial derivatives $(\partial f_M)/(\partial x_{t-j})(\mathbf{x})$ for $j = 1,\ldots,p$ can be obtained by minimizing, $\sum_{t=1}^{T-M+1}(x_{t+M-1} - \beta_{M,0} - \beta'_{M,1}(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1}-\mathbf{x})$, or $\widehat{\beta}_{M,0}(\mathbf{x}) = \widehat{f}_M(\mathbf{x})$ and

$$\widehat{\beta}_{M,1}(\mathbf{x}) = \begin{bmatrix} \widehat{\beta}_{M,11}(\mathbf{x}) \\ \vdots \\ \widehat{\beta}_{M,1p}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \Delta\widehat{f}_{M,1}(\mathbf{x}) \\ \vdots \\ \Delta\widehat{f}_{M,p}(\mathbf{x}) \end{bmatrix} ,$$

respectively, where $\widehat{\beta}_M(x) = (\widehat{\beta}_{M,0}(\mathbf{x}), \widehat{\beta}_{M,1}(\mathbf{x})')' = (\mathsf{X}_x'\mathbf{W}_x\mathsf{X}_x)^{-1}\mathsf{X}_x'\mathbf{W}_x\mathbf{Y}$,

$$\mathsf{X}_x = \begin{bmatrix} 1 & (\mathsf{X}_0-\mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathsf{X}_{T-M}-\mathbf{x})' \end{bmatrix},$$

$\mathbf{Y} = (x_M,\ldots,x_T)'$ and $W_x = \mathrm{diag}\,\{K_H(\mathsf{X}_0-\mathbf{x}),\ldots, K_H(\mathsf{X}_{T-M}-\mathbf{x})\}$. The local linear estimator of the IRF is then given by

$$\widehat{\mathrm{IRF}}_M(\delta,\mathbf{x}) = \widehat{f}_M(\mathbf{x}^*) - \widehat{f}_M(\mathbf{x})$$

where $\mathbf{x}^* = (\overline{x}_0+\delta, \overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$. Similarly, the estimator of the alternative IRF is given by

$$\widehat{\mathrm{IRF}}^*_M(\delta,\mathbf{x}) = \widehat{f}_{M-1}(\mathbf{x}^*) - \widehat{f}_M(\mathbf{x})$$

where $\mathbf{x}^* = (\widehat{f}(\mathbf{x})+\delta, \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$. When $\mathbf{x}$ and δ are given, computing nonparametric IRFs for a sequence of M provide a useful information on the persistence of deviation without specifying the autoregressive function. However, instead of reporting IRFs for many possible combinations of $\mathbf{x}$ and δ, one can also compute the small deviation IRF based on the nonparametric estimate of the first partial derivative at $\mathbf{x}$. The local linear estimator of the small deviation IRF is given by $\Delta\widehat{f}_{M,1}(\mathbf{x})$ for the first version, and $\Delta\widehat{f}_{M-1,1}(\widehat{f}(\mathbf{x}), \overline{x}_0, \overline{x}_{-1},\ldots,\overline{x}_{-p+2})$ for the second version, respectively. A large change in the value of derivatives with increasing M represents the sensitive dependence on initial conditions.

Yao and Tong's Variance Decomposition

The initial value sensitivity of the system with dynamic noise also has an implication in the presence of additional observation noise. Suppose that current observation is subject to a measurement error, a rounding error, or when only preliminary estimates of aggregate economic variables announced by the statistical agency are available. When the true current position deviates slightly from $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$, the performance of the same predictor may be measured by $E[\{x_{t+M} - f_M(\mathbf{x})\}^2 | \mathsf{X}_t = \mathbf{x} + \boldsymbol{\delta}]$. Under a certain condition, this MSFE can be decomposed as follows:

$$\begin{aligned} &E\left[\{x_{t+M} - f_M(\mathbf{x})\}^2 | \mathsf{X}_t = \mathbf{x} + \boldsymbol{\delta}\right] \\ &\quad = \sigma_M^2(\mathbf{x} + \boldsymbol{\delta}) + \{f_M(\mathbf{x} + \boldsymbol{\delta}) - f_M(\mathbf{x})\}^2 \\ &\quad = \sigma_M^2(\mathbf{x} + \boldsymbol{\delta}) + \{\boldsymbol{\delta}' \Delta f_M(\mathbf{x})\}^2 + o(||\boldsymbol{\delta}||^2) \end{aligned}$$

where $\Delta f_M(\mathbf{x}) = (\Delta f_{M,1}(\mathbf{x}), \ldots, \Delta f_{M,p}(\mathbf{x}))'$, $\Delta f_{M,j}(\mathbf{x}) = (\partial f_M)/(\partial x_{t-j})(\mathbf{x})$ for $j = 1, \ldots, p$. This decomposition shows two dominant components in the MSFE. The first component represents the prediction error caused by the randomness in the system at point $\mathbf{x} + \boldsymbol{\delta}$. This component will be absent in the case where there is no dynamic noise ε_t in the system. The second component represents the difference caused by the deviation $\boldsymbol{\delta}$ from the initial point $\mathbf{x}$. When the non-zero deviation $\boldsymbol{\delta}$ is much smaller than σ, the standard deviation of ε_t, the first component $\sigma_M^2(\mathbf{x} + \boldsymbol{\delta}) = \sigma_M^2(\mathbf{x}) + O(||\boldsymbol{\delta}||)$ is the dominant term because the second component $\{\boldsymbol{\delta}' \Delta f_M(\mathbf{x})\}^2$ is of order $O(||\boldsymbol{\delta}||^2)$. However, for a nonlinear system with a very small error ε_t, the contribution of the second term can become nonnegligible. Thus, [80] considered $\Delta f_M(\mathbf{x})$ as a measure of sensitivity to initial conditions for the M-period ahead forecast (They referred to the M-step Lyapunov-like index).

If $f_M(\mathbf{x})$ is replaced by a mean square consistent estimator $\widehat{f}_M(\mathbf{x})$, such as a local linear estimator $\widehat{\beta}_{M,0}(\mathbf{x})$,

$$\begin{aligned} &\lim_{T\to\infty} E\left[\{x_{T+M} - \widehat{f}_M(\mathbf{x})\}^2 | \mathsf{X}_T = \mathbf{x} + \boldsymbol{\delta}\right] \\ &\qquad = \sigma_M^2(\mathbf{x} + \boldsymbol{\delta}) + \{\boldsymbol{\delta}' \Delta f_M(\mathbf{x})\}^2 + o(||\boldsymbol{\delta}||^2) \,. \end{aligned}$$

Thus the decomposition is still valid. For the estimation of the sensitivity measure $\Delta f_M(\mathbf{x})$, the local linear estimator $\widehat{\beta}_{M,0}(\mathbf{x}) = \Delta \widehat{f}_M(\mathbf{x})$ can be used. In practice, it is convenient to consider a norm version of the measure $\mathrm{LI}_M(\mathbf{x}) = ||\Delta f_M(\mathbf{x})||$ and report its estimator

$$\widehat{\mathrm{LI}}_M(\mathbf{x}) = ||\Delta \widehat{f}_M(\mathbf{x})||$$

evaluated at various $\mathbf{x}$. In a one-dimensional case, they are $\mathrm{LI}_M(\overline{x}_0) = |f_M'(\overline{x}_0)|$ and $\widehat{\mathrm{LI}}_M(\overline{x}_0) = |\widehat{f'}_M(\overline{x}_0)|$, respectively. Note that $\mathrm{LI}_M(\mathbf{x})$ is related to the derivative of the normalized nonlinear impulse response function $\mathrm{IRF}_M(\delta, \mathbf{x})$. Recall that, in the one-dimensional case, a normalized IRF to infinitesimal shocks becomes the first derivative. Thus, $\mathrm{LI}_M(\overline{x}_0)$ is the absolute value of the estimator of the corresponding IRF. In the multidimensional case, IRF to small shocks becomes the partial derivative with respect to the first components. If shocks are also given to other initial values in IRF, computing the norm of the estimator of all IRFs yields $\mathrm{LI}_M(\mathbf{x})$.

This sensitivity measure is also related to the local Lyapunov exponent. In the one-dimensional case, with a fixed M, the local Lyapunov exponent can be written as $\lambda_M(\overline{x}_0) = M^{-1} \ln |\prod_{t=1}^{M} f'(x_{t-1})|$. If the contribution of ε_t is very small, $\mathrm{d}f^{(M)}(x_0)/\mathrm{d}x \approx \prod_{t=1}^{M} f'(x_{t-1})$ and then the estimator $\widehat{f'}_M(x_0)$ becomes an estimator of $\mathrm{d}f^{(M)}(x_0)/\mathrm{d}x$. Thus $\lambda_M(\overline{x}_0)$ can be also estimated by $M^{-1} \ln \widehat{\mathrm{LI}}_M(\overline{x}_0)$.

Information Matrix

The last measure of the initial value sensitivity is the one based on the distance between two distributions of M-steps ahead forecast, conditional on two nearby initial values $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ and $\mathbf{x} + \boldsymbol{\delta}$ where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$. Let $\pi_M(y|\mathbf{x})$ and $\Delta\pi_M(y|\mathbf{x})$ be the conditional density function of x_M given $\mathsf{X}_0 = \mathbf{x}$ and a $p \times 1$ vector of its partial derivatives. [81] suggested using Kullback–Leibler information to measure the distance, which is given by

$$\begin{aligned} K_M(\delta, \mathbf{x}) = \int_{-\infty}^{+\infty} &\left\{\pi_M(y|\mathbf{x} + \boldsymbol{\delta}) - \pi_M(y|\mathbf{x})\right\} \\ &\cdot \ln\left\{\pi_M(y|\mathbf{x} + \boldsymbol{\delta}) / \pi_M(y|\mathbf{x})\right\} \mathrm{d}y \,. \end{aligned}$$

Assuming the smoothness of conditional distribution and interchangeability of integration and differentiation, Taylor series expansion around $\mathbf{x}$ for small $\boldsymbol{\delta}$ yields

$$K_M(\delta, \mathbf{x}) = \boldsymbol{\delta}' I_M(\mathbf{x}) \boldsymbol{\delta} + o(||\boldsymbol{\delta}||^2)$$

where

$$I_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\pi_M(y|\mathbf{x}) \Delta\pi_M(y|\mathbf{x})' / \pi_M(y|\mathbf{x}) \mathrm{d}y \,.$$

If initial value $\mathbf{x}$ is treated as a parameter vector of the distribution, $I_M(\mathbf{x})$ is the Fisher's information matrix, which represents the information on $\mathbf{x}$ contained in x_M. This quantity can be used as an initial value sensitivity measure since more information on $\mathbf{x}$ implies more sensitivity of distribution of x_M to the initial condition $\mathbf{x}$. This information matrix measure and the M-step Lyapunov-like index are related via the following inequality when the system is one-dimensional,

$$I_M(\overline{x}_0) \geq \frac{\mathrm{LI}_M^2(\overline{x}_0)}{\sigma_M^2(\overline{x}_0)} \,.$$

Thus, for a given M-step Lyapunov-like index, a larger conditional MSFE implies more sensitivity. In addition, because that $\lambda_M(\overline{x}_0) \approx M^{-1} \ln \mathrm{LI}_M(\overline{x}_0)$ and $\sigma_M^2(\overline{x}_0) \approx \sigma^2[1 + \sum_{j=1}^{M-1} \exp\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\}]$,

$$\ln I_M(\overline{x}_0) \geq 2M\lambda_M(\overline{x}_0) \\ -\ln\left[1 + \sum_{j=1}^{M-1} \exp\left\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\right\}\right] - \ln \sigma^2$$

holds approximately.

As an alternative to Kullback–Leibler distance, [28] considered L_2-distance given by

$$D_M(\delta, \mathbf{x}) = \int_{-\infty}^{+\infty} \{\pi_M(y|\mathbf{x} + \boldsymbol{\delta}) - \pi_M(y|\mathbf{x})\}^2 \mathrm{d}y \, .$$

Because of a similar argument, for small $\boldsymbol{\delta}$, $D_M(\delta, \mathbf{x})$ can be approximated by

$$D_M(\delta, \mathbf{x}) = \boldsymbol{\delta}' J_M(\mathbf{x})\boldsymbol{\delta} + o(||\boldsymbol{\delta}||^2)$$

where

$$J_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\pi_M(y|\mathbf{x})\Delta\pi_M(y|\mathbf{x})' \mathrm{d}y \, .$$

Note that $J_M(\mathbf{x})$ cannot be interpreted as Fisher's information but can still be used as a sensitivity measure.

Both $I_M(\mathbf{x})$ and $J_M(\mathbf{x})$ can be estimated nonparametrically. Consider the minimization problem of $\sum_{t=1}^{T-M+1}(\kappa_h(x_{t+M-1} - y) - \beta_{M,0} - \beta'_{M,1}(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$ where $\kappa_h(u) = \kappa(u/h)/h$, h is the bandwidth and κ is a univariate kernel function, instead of minimizing $\sum_{t=1}^{T-M+1}(x_{t+M-1} - \beta_{M,0} - \beta'_{M,1}(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$. Then, $\widehat{\beta}_{M,0}(\mathbf{x}, y) = \widehat{\pi}_M(y|\mathbf{x})$ and $\widehat{\beta}_{M,1}(\mathbf{x}, y) = \Delta\widehat{\pi}_M(y|\mathbf{x})$, where $\widehat{\beta}_M(\mathbf{x}, y) = (\widehat{\beta}_{M,0}(\mathbf{x}, y), \widehat{\beta}_{M,1}(\mathbf{x}, y)')' = (\mathsf{X}'_x \mathbf{W}_x \mathsf{X}_x)^{-1}\mathsf{X}'_x \mathbf{W}_x \mathbf{Y}_y$,

$$\mathsf{X}_x = \begin{bmatrix} 1 & (\mathsf{X}_0 - \mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathsf{X}_{T-M} - \mathbf{x})' \end{bmatrix} ,$$

$\mathbf{Y}_y = \{\kappa_h(x_M - y), \ldots, \kappa_h(x_T - y)\}'$ and $W_x = \mathrm{diag}\,\{K_H(\mathsf{X}_0 - \mathbf{x}), \ldots, K_H(\mathsf{X}_{T-M} - \mathbf{x})\}$. Then the estimators of $I_M(\mathbf{x})$ and $J_M(\mathbf{x})$ are given by

$$\widehat{I}_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\widehat{\pi}_M(y|\mathbf{x})\Delta\widehat{\pi}_M(y|\mathbf{x})'/\widehat{\pi}_M(y|\mathbf{x})\mathrm{d}y$$

and

$$\widehat{J}_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\widehat{\pi}_M(y|\mathbf{x})\Delta\widehat{\pi}_M(y|\mathbf{x})' \mathrm{d}y \, ,$$

respectively.

Forecasting Financial Asset Returns and Sensitive Dependence

Nonlinear Forecasting of Asset Returns

In this subsection, a quick review of the general issues of forecasting financial asset returns is first provided, then the empirical results on the nonlinear forecasting based on nonparametric methods are summarized.

In the past, the random walk model was considered as the most appropriate model to describe the dynamics of asset prices in practice (see [24]). However, after decades of investigation, more evidence on some predictable components of asset returns has been documented in the literature. Although the evidence is often not very strong, several studies report the positive serial dependence for relatively short horizon stock returns. For example, [47] show that first-order autocorrelation of weekly returns on the Center for Research in Security Prices (CRSP) index is as high as 30 percent and significant when an equal-weighted index is used, but is somewhat less when a value-weighted index is used ([12] provide similar evidence for the daily return). The conditional mean of stock returns may not depend only on the past returns but also on other economic variables, including dividend yields, price earnings ratio, short and long interest rates, industrial production and inflation rate. A comprehensive statistical analysis to evaluate the 1-month-ahead out-of-sample forecast of 1 month excess returns by these predictors is conducted by [55]. Some results on the long-horizon predictability in stock returns, based on lagged returns (e. g., [25] and [57]) and other economic variables such as dividend yields or dividend-price ratios (e. g., [26] and [13]) are also available. This evidence on long-horizon forecasts, however, is still controversial because the standard statistical inference procedure may not be reliable in case when the correlation coefficient is computed from a small number of nonoverlapping observations [62] or when the predictor is very persistent in the forecasting regression [73].

The question is whether the introduction of nonlinear structure helps improve the forecasting performance of future asset returns. When the nonlinear condition mean function is unspecified, the neural network method has often been employed as a reliable nonparametric method in predicting the returns. For IBM daily stock returns, [75] found no improvement in out-of-sample predictability based on the neural network model. For daily returns of the Dow Jones Industrial Average (DJIA) index, [33] estimated a nonlinear AR model using the same method. He, in contrast, showed that MSFE reduction over a benchmark linear AR model could be as large as 12.3 percent for the 10-day-ahead out-of-sample forecast. The role of economic fundamentals as predictors can be also investigated under the nonlinear framework. Using a model selection procedure similar to the one employed by [55], some evidence of MSFE improvement from neural network-based forecast of excess returns was provided in [60] and [59] but no encouraging evidence was found in similar studies by [61] and [49]. In practice, 'noise traders' or 'chartists' may predict prices using some technical trading rules (TTRs) rather than using economic fundamentals. For example, a simple TTR based on the moving average can generate a buy signal when the current asset price level P_t is above $n^{-1}\sum_{i=1}^{n} P_{t-i+1}$ for some positive integer n and a sell signal when it is below. [11] found some evidence on the nonlinearity in the conditional mean of DJIA returns conditional on buy-sell signals. [33] further considered including past buy-sell signals as predictors in the neural network model and found that an improvement in MSFE over the linear AR model was even larger than the case when only lagged returns are used as a predictor in the neural network model. One useful nonlinear model is the functional coefficient AR model where the AR coefficient can depend on time or some variables. For example, as in [39], the AR coefficient can be a function of buy-sell signals. [44] claimed that a functional coefficient AR model with a coefficient as a function of the moving average of squared returns well described the serial correlation feature of stock returns.

This moderate but increasing evidence of nonlinear forecastability applies not only to the stock market but also to the foreign exchange market. In the past, [52] could not find any reasonable linear model that could out-perform the random walk model in an out-of-sample forecast of foreign exchange rates. The nonlinear AR model was estimated nonparametrically by [19] but no forecasting improvement over the random walk model could be found in their analysis. However, many follow-up studies, including [34,39,43,76], provided some evidence on forecastability with nonlinear AR models estimated using neural networks or other nonparametric methods.

One important and robust empirical fact is that much higher positive serial correlation is typically observed for the volatility measures such as the absolute returns, $|x_t|$, and their power transformation, $|x_t|^{\alpha}$ for $\alpha > 0$, than for the returns, x_t ([20,69]). This observation is often referred to as a volatility clustering. As a result, forecasting volatility has been much more successful than forecasting returns themselves. The most commonly used approach in forecasting volatility is to describe the conditional variance of asset returns using the class of ARCH and GARCH models ([8,23]). The volatility of stock returns is also known to respond more strongly to negative shocks in returns than positive ones. This 'leverage effect' often motivates the introduction of nonlinear structure in volatility modeling such as the EGARCH model of [53]. Instead of estimating the unknown parameter in a specified ARCH model, the nonparametric method can be also employed to estimate the possibly nonlinear ARCH model in forecasting (see [46]). The better forecastability of market direction (or market timing), sign (x_t), than that of returns, has also been documented in the literature. Examples are [55] for the stock market and [39,43], and [16] for the foreign exchange market. Since the return, x_t, can be decomposed into a product of the two components, $|x_t| \times \text{sign}\,(x_t)$, one may think the strong linear or nonlinear forecastability of the volatility and the sign of returns should lead to forecastability of the returns as well. Interestingly, however, [17] theoretically showed that the serial dependence of asset return volatilities and that of return signs did not necessarily imply the serial dependence of returns.

In summary, a growing number of recent studies show some evidence of linear and nonlinear forecastability of asset returns, and stronger evidence of forecastability of their nonlinear transformations, such as the squared returns, absolute returns and the sign of returns. In this sense, the nonlinearity seems to be playing a non-negligible role in explaining the dynamic behavior of asset prices.

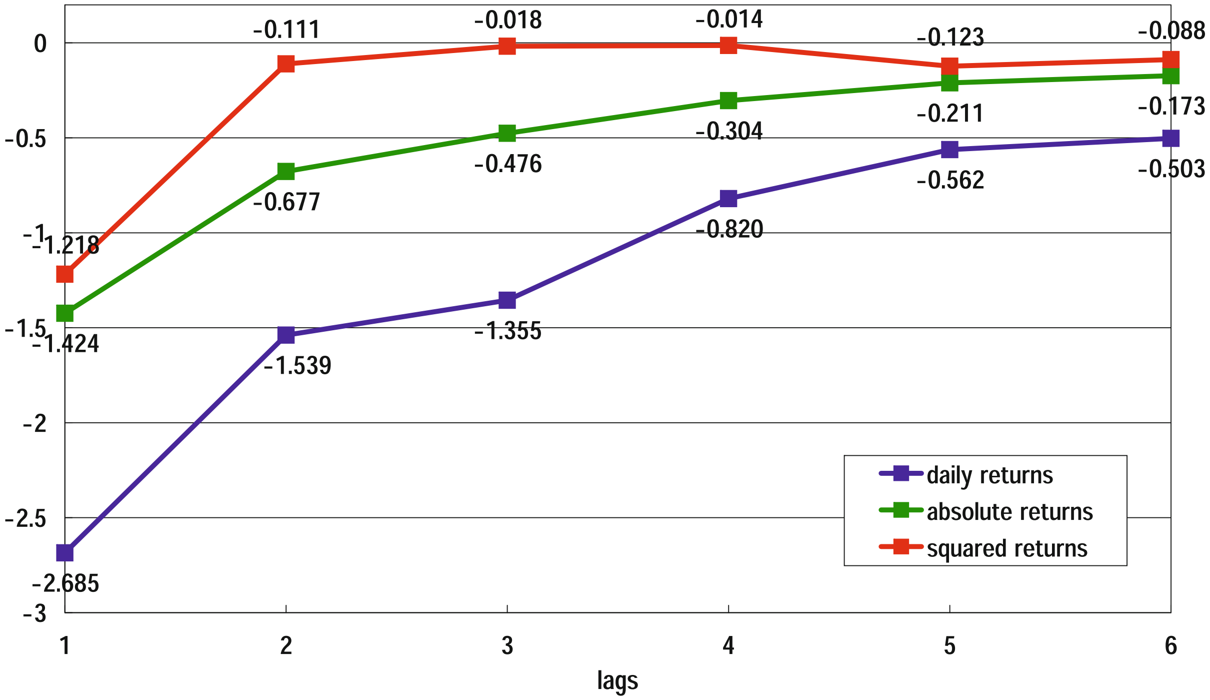
Initial Value Sensitivity in Financial Data

Theoretically, when investors have heterogeneous expectations about the future prices, asset price dynamics can be chaotic with a positive Lyapunov exponent [10]. A comprehensive list on earlier empirical work related to the sensitive dependence and chaos in financial data is provided in [2,6]. Many early studies employed either the BDS test or a dimension estimator and provided the indirect evidence on sensitive dependence and chaos. For example, [64] applied the BDS test to weekly returns on the value-weighted CRSP portfolio and rejected iid randomness. [41] further examined weekly value-weighted and equally weighted CRSP portfolio returns, as well as Standard & Poor 500 (S&P 500) index returns for various frequencies, and found strong evidence against iid. Similar findings are also reported for the daily foreign exchange rate returns in [40]. For financial variables, high-frequency data or tick data is often available to researchers. Earlier examples of studies on chaos using high-frequency data include [50], who found some evidence of low-dimensional chaos based on the correlation dimension and K_2 entropy of 20-second S&P 500 index returns, with a number of observations as large as 19,027. Estimation results on Lyapunov exponents for high-frequency stock returns are also available. In addition to the BDS test, [1] and [2] employed the neural network method and found negative Lyapunov exponents in 1- and 5-minute returns of cash series of S&P 500, UK Financial Times Stock Exchange-100 (FTSE-100) index, Deutscher Aktienindex (DAX), the Nikkei 225 Stock Average, and of futures series of S&P 500 and FTSE-100. Using the resampling procedure of [32], [65] obtained significantly negative Lyapunov exponents for daily stock returns of the Austrian Traded Index (ATX). For the foreign exchange market, [18] estimated Lyapunov exponents of the Canadian, German, Italian and Japanese monthly spot exchange rates using neural nets and found some mixed result regarding their sign.

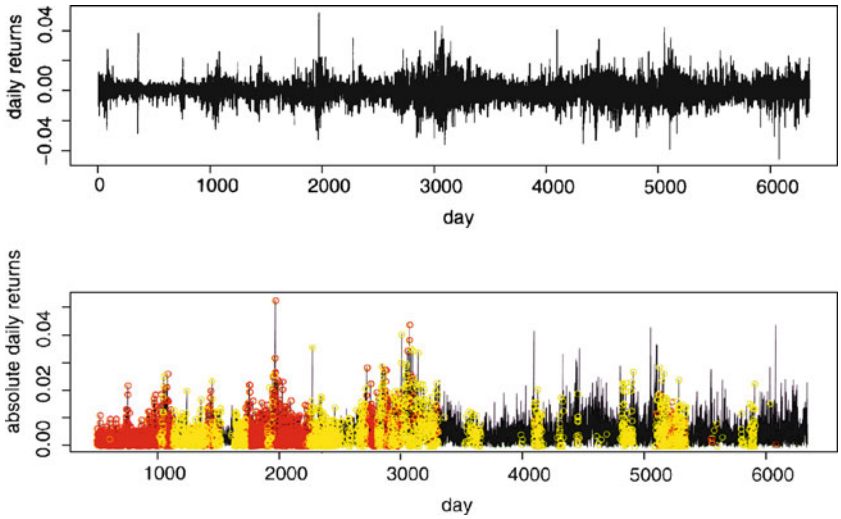
By using the absolute returns or their power transformation instead of using returns themselves, sensitive dependence of volatility on initial conditions may be examined nonparametrically. [68] used neural nets and estimated Lyapunov exponents of higher order daily returns of the DJIA index. Figure 5 shows their global Lyapunov exponent estimates for simple returns, squared returns and absolute returns. For all cases, Lyapunov exponents are significantly negative but the values of absolute returns are always larger than that of simple returns. While some estimates are close to zero, the observation of the monotonically increasing Lyapunov exponent with increasing *p*, for daily and absolute returns, resembles the simulation results of the previous section implying the upward bias when the true Lyapunov exponent is negative.

For the exchange rate market, [29] applied [63]'s method to absolute changes and their power transformation of Canadian and German nominal exchange rates and did not reject the null hypothesis of chaos.

For a local measure of initial value sensitivity, [68] also reported the median values of 145 estimates of local Lyapunov exponents for DJIA returns, in addition to the global Lyapunov exponents. [45] reported the nonlinear impulse response functions of yen/dollar and deutschemark/dollar exchange rate returns based on parametrically estimated GARCH model. [14] reported a Lyapunov-like index of [80] for the simple returns and absolute returns of CRSP data used in [64]. From Fig. 6, they concluded that (i) the first half of the CRSP series is more volatile than its second half, suggesting that the market becomes more mature with time, and (ii) volatile periods tend to form clusters. [65] reported the information matrix measure of local sensitive dependence computed from ATX data based on the parametric estimation of ARCH



Financial Forecasting, Sensitive Dependence, Figure 5
Global Lyapunov exponents of stock returns



Financial Forecasting, Sensitive Dependence, Figure 6
Upper **panel displays the time series plot of the CRSP daily returns.** ***Lower*** **panel shows the absolute CRSP daily returns with data coloured** ***red*** **whenever their Lyapunov-like indices are above the third quartile of the indices, and data coloured** ***yellow*** **if their indices are between the median and the third quartile**

and GARCH models, in addition to nonparametric estimates of the global Lyapunov exponent.

On the whole, empirical studies on global and local sensitivity measures suggested less sensitive dependence than the chaotic model would predict, but some sensitivity of short-term forecastability on initial conditions.

Future Directions

Some of the possible directions of future research topics are in order.

The first direction is to search for the economic theory behind the initial value sensitivity if detected in the data. The statistical procedures introduced here are basically data description and the empirical results obtained by this approach are not directly connected to underlying economic or finance theory. Theories, such as the one developed by [10], can predict complex behavior of asset prices but direct estimation of the model are typically not possible. Thus, for most cases, the model is evaluated by matching the actual data with the one generated from the model in simulation. Thus direct implication to the sensitive dependence measure would provide a more convincing argument for the importance of knowing the structure. [38] may be considered as one attempt in this direction.

The second direction is to develop better procedures in estimating the initial value sensitivity with the improved accuracy in the environment of a relatively small sample size. In the Jacobian method of estimating the Lyapunov exponent, the conditional mean function has been estimated either parametrically and nonparametrically. A fully nonparametric approach, however, is known to suffer from a high dimensionality problem. A semiparametric approach, such as the one for an additive AR model, is likely to be useful in this context but has not been used in the initial value sensitivity estimation.

The third direction is towards further analysis based on high-frequency data, which has become more commonly available in empirical finance. Much progress has been made in the statistical theory on the realized volatility computed from such data, and forecasting volatility of asset returns based on the realized volatility has been empirically successful (see, e. g., [3]). However, so far, this approach has not been used in detecting the initial value sensitivity in volatility. In addition, realized volatility is known to suffer from market microstructure noise when sampling frequency increases. Given the fact that the initial value sensitivity measures can be considered in the framework of the nonlinear AR models, namely, the stochastic environment in the presence of noise, it is of interest in investigating the robustness of the procedure to the market microstructure noise when applied to high-frequency returns.

Bibliography

1. Abhyankar A, Copeland LS, Wong W (1995) Nonlinear dynamics in real-time equity market indices: evidence from the United Kingdom. Econ J 105:864–880

2. Abhyankar A, Copeland LS, Wong W (1997) Uncovering nonlinear structure in real-time stock-market indexes: The S&P 500, the DAX, the Nikkei 225, and the FTSE-100. J Bus Econ Stat 15:1–14
3. Andersen TB, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71:579–625
4. Bailey BA, Ellner S, Nychka DW (1997) Chaos with confidence: Asymptotics and applications of local Lyapunov exponents. In: Cutler CD, Kaplan DT (eds) Fields Institute Communications, vol 11. American Mathematical Society, Providence, pp 115–133
5. Barnett WA, Gallant AR, Hinich MJ, Jungeilges J, Kaplan D, Jensen MJ (1995) Robustness of nonlinearity and chaos tests to measurement error, inference method, and sample size. J Econ Behav Organ 27:301–320
6. Barnett WA, Serletis A (2000) Martingales, nonlinearity, and chaos. J Econ Dyn Control 24:703–724
7. Bask M, de Luna X (2002) Characterizing the degree of stability of non-linear dynamic models. Stud Nonlinear Dyn Econom 6:3
8. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econ 31:307–327
9. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econ Rev 15(3):197–235
10. Brock WA, Hommes CH (1998) Heterogeneous beliefs and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
11. Brock WA, Lakonishok J, LeBaron B (1992) Simple technical trading rules and the stochastic properties of stock returns. J Financ 47:1731–1764
12. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
13. Campbell JY, Shiller R (1988) The dividend-price ratio and expectations of future dividends and discount factors. Rev Financ Stud 1:195–228
14. Chan KS, Tong H (2001) Chaos: A Statistical Perspective. Springer, New York
15. Cheng B, Tong H (1992) On consistent nonparametric order determination and chaos. J Royal Stat Soc B 54:427–449
16. Cheung YW, Chinn MD, Pascual AG (2005) Empirical exchange rate models of the nineties: Are any fit to survive? J Int Money Financ 24:1150–1175
17. Christoffersen PF, Diebold FX (2006) Financial asset returns, direction-of-change forecasting, and volatility dynamics. Management Sci 52:1273–1287
18. Dechert WD, Gençay R (1992) Lyapunov exponents as a nonparametric diagnostic for stability analysis. J Appl Econom 7:S41–S60
19. Diebold FX, Nason JA (1990) Nonparametric exchange rate prediction? J Int Econ 28:315–332
20. Ding Z, Granger CWJ, Engle RF (1993) A long memory property of stock market returns and a new model. J Empir Financ 1: 83–106
21. Eckmann JP, Kamphorst SO, Ruelle D, Ciliberto S (1986) Liapunov exponents from time series. Phys Rev A 34:4971–4979
22. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
23. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50:987–1008
24. Fama E (1970) Efficient capital markets: Review of theory and empirical work. J Financ 25:383–417
25. Fama E, French K (1988) Permanent and temporary components of stock prices. J Political Econ 96:246–273
26. Fama E, French K (1988) Dividend yields and expected stock returns. J Financ Econ 22:3–5
27. Fan J, Yao Q (2003) Nonlinear Time Series: Nonparametric and Parametric Methods. Springer, New York
28. Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83:189–206
29. Fernándes-Rodríguez F, Sosvilla-Rivero S, Andrada-Félix J (2005) Testing chaotic dynamics via Lyapunov exponents. J Appl Econom 20:911–930
30. Frank M, Stengos T (1989) Measuring the strangeness of gold and silver rates of return. Rev Econ Stud 56:553–567
31. Gallant AR, Rossi PE, Tauchen G (1993) Nonlinear dynamic structures. Econometrica 61:871–907
32. Gençay R (1996) A statistical framework for testing chaotic dynamics via Lyapunov exponents. Physica D 89:261–266
33. Gençay R (1998) The predictability of security returns with simple technical trading rules. J Empir Financ 5:347–359
34. Gençay R (1999) Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. J Int Econ 47:91–107
35. Giannerini S, Rosa R (2001) New resampling method to assess the accuracy of the maximal Lyapunov exponent estimation. Physica D 155:101–111
36. Grassberger P, Procaccia I (1983) Estimation of the Kolmogorov entropy from a chaotic signal. Phys Rev A 28:2591–2593
37. Hall P, Wolff RCL (1995) Properties of invariant distributions and Lyapunov exponents for chaotic logistic maps. J Royal Stat Soc B 57:439–452
38. Hommes CH, Manzan S (2006) Comments on Testing for nonlinear structure and chaos in economic time series. J Macroecon 28:169–174
39. Hong Y, Lee TH (2003) Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. Rev Econ Stat 85:1048–1062
40. Hsieh DA (1989) Testing for nonlinear dependence in daily foreign exchange rates. J Bus 62:339–368
41. Hsieh DA (1991) Chaos and nonlinear dynamics: application to financial markets. J Financ 46:1839–1877
42. Koop G, Pesaran MH, Potter SM (1996) Impulse response analysis in nonlinear multivariate models. J Econom 74:119–147
43. Kuan CM, Liu T (1995) Forecasting exchange rates using feedforward and recurrent neural networks. J Appl Econom 10:347–364
44. LeBaron B (1992) Some relation between the volatility and serial correlations in stock market returns. J Bus 65:199–219
45. Lin WL (1997) Impulse response function for conditional volatility in GARCH models. J Bus Econ Stat 15:15–25
46. Linton OB (2008) Semiparametric and nonparametric ARCH modelling. In: Anderson TG, Davis RA, Kreiss JP, Mikosch T (ed) Handbook of Financial Time Series. Springer, Berlin
47. Lo AW, MacKinlay AC (1988) Stock market prices do not follow random walks: evidence from a simple specification test. Rev Financ Stud 1:41–66
48. Lu ZQ, Smith RL (1997) Estimating local Lyapunov exponents. In: Cutler CD, Kaplan DT (eds) Fields Institute Communica-

tions, vol 11. American Mathematical Society, Providence, pp 135–151

49. Maasoumi E, Racine J (2002) Entropy and predictability of stock market returns. J Econom 107:291–312
50. Mayfield ES, Mizrach B (1992) On determining the dimension of real time stock price data. J Bus Econ Stat 10:367–374
51. McCaffrey DF, Ellner S, Gallant AR, Nychka DW (1992) Estimating the Lyapunov exponent of a chaotic system with nonparametric regression. J Am Stat Assoc 87:682–695
52. Meese R, Rogoff K (1983) Exchange rate models of the seventies. Do they fit out of sample? J Int Econ 14:3–24
53. Nelson DB (1990) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
54. Nychka D, Ellner S, Gallant AR, McCaffrey D (1992) Finding chaos in noisy system. J Royal Stat Soc B 54:399–426
55. Pesaran MH, Timmermann A (1995) Predictability of stock returns: robustness and economic significance. J Financ 50: 1201–1228
56. Pesin JB (1977) Characteristic Liapunov exponents and smooth ergodic theory. Russ Math Surv 32:55–114
57. Poterba JM, Summers LH (1988) Mean reversion in stock prices: evidence and implications. J Financ Econ 22:27–59
58. Potter SM (2000) Nonlinear impulse response functions. J Econ Dyn Control 24:1425–1446
59. Qi M (1999) Nonlinear predictability of stock returns using financial and economic variables. J Bus Econ Stat 17:419–429
60. Qi M, Maddala GS (1999) Economic factors and the stock market: a new perspective. J Forecast 18:151–166
61. Racine J (2001) On the nonlinear predictability of stock returns using financial and economic variables. J Bus Econ Stat 19: 380–382
62. Richardson M, Stock JH (1989) Drawing inferences from statistics based on multiyear asset returns. J Financ Econ 25: 323–348
63. Rosenstein MT, Collins JJ, De Luca CJ (1993) A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 65:117–134
64. Scheinkman JA, LeBaron B (1989) Nonlinear dynamics and stock returns. J Bus 62:311–337
65. Schittenkopf C, Dorffner G, Dockner EJ (2000) On nonlinear, stochastic dynamics in economic and financial time series. Stud Nonlinear Dyn Econom 4:101–121
66. Shintani M (2006) A nonparametric measure of convergence towards purchasing power parity. J Appl Econom 21:589–604
67. Shintani M, Linton O (2003) Is there chaos in the world economy? A nonparametric test using consistent standard errors. Int Econ Rev 44:331–358
68. Shintani M, Linton O (2004) Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. J Econom 120:1–33
69. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
70. Tjostheim D, Auestad BH (1994) Nonparametric identification of nonlinear time series: Selecting significant lags. J Am Stat Assoc 89:1410–1419
71. Tschernig R, Yang L (2000) Nonparametric lag selection for time series. J Time Ser Analysis 21:457–487
72. Tschernig R, Yang L (2000) Nonparametric estimation of generalized impulse response functions. Michigan State University, unpublished
73. Valkanov R (2003) Long-horizon regressions: theoretical results and applications. J Financ Econom 68:201–232
74. Whang YJ, Linton O (1999) The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic time series. J Econom 91:1–42
75. White H (1988) Economic prediction using neural networks: the case of IBM stock returns. Proceedings of the IEEE International Conference on Neural Networks 2. The Institute of Electrical and Electronics Engineers, San Diego, pp 451–458
76. White H, Racine J (2001) Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. IEEE Trans Neural Netw 12:657–673
77. Wolf A, Swift JB, Swinney HL, Vastano JA (1985) Determining Lyapunov exponents from a time series. Physica D 16:285–317
78. Wolff RCL (1992) Local Lyapunov exponent: Looking closely at chaos. J Royal Stat Soc B 54:353–371
79. Wolff R, Yao Q, Tong H (2004) Statistical tests for Lyapunov exponents of deterministic systems. Stud Nonlinear Dyn Econom 8:10
80. Yao Q, Tong H (1994) Quantifying the influence of initial values on non-linear prediction. J Royal Stat Soc Ser B 56:701–725
81. Yao Q, Tong H (1994) On prediction and chaos in stochastic systems. Philos Trans Royal Soc Lond A 348:357–369

Fractals and Economics

Misako Takayasu[1], Hideki Takayasu[2]
[1] Tokyo Institute of Technology, Tokyo, Japan
[2] Sony Computer Science Laboratories Inc, Tokyo, Japan

Article Outline

Glossary

Fractal An adjective or a noun representing complex configurations having scale-free characteristics or self-similar properties. Mathematically, any fractal can be characterized by a power law distribution.

Power law distribution For this distribution the probability density is given by a power law, $p(r) = c \cdot r^{-\alpha-1}$, where c and α are positive constants.

Foreign exchange market A free market of currencies, exchanging money in one currency for other, such as purchasing a United States dollar (USD) with Japanese yen (JPY). The major banks of the world are trading 24 hours and it is the largest market in the world.

Definition of the Subject

Market price fluctuation was the very first example of fractals, and since then many examples of fractals have been found in the field of Economics. Fractals are everywhere in economics. In this article the main attention is focused on real world examples of fractals in the field of economics, especially market properties, income distributions, money flow, sales data and network structures. Basic mathematics and physics models of power law distributions are reviewed so that readers can start reading without any special knowledge.

Introduction

Fractal is the scientific word coined by B.B. Mandelbrot in 1975 from the Latin word *fractus*, meaning "fractured" [25]. However, *fractal* does not directly mean fracture itself. As an image of a fractal Fig. 1 shows a photo of fractured pieces of plaster fallen on a hard floor. There are several large pieces, many middle size pieces and countless fine pieces. If you have a microscope and observe a part of floor carefully then you will find in your vision several large pieces, many small pieces and countless fine pieces, again in the microscopic world. Such scale-invariant nature is the heart of the fractal. There is no explicit definition on the word fractal, it generally means a complicated scale-invariant configuration.

Fractals and Economics, Figure 1
Fractured pieces of plaster fallen on a hard floor (provided by H. Inaoka)

Scale-invariance can be defined mathematically [42]. Let $P(\geq r)$ denote the probability that the diameter of a randomly chosen fractured piece is larger than r, then this distribution is called scale-invariant if this function satisfies the following proportional relation for any positive scale factor λ in a considering scale range:

$$P(\geq \lambda r) \propto P(\geq r)\,. \tag{1}$$

The proportional factor should be a function of λ, so we can re-write Eq. (1) as

$$P(\geq \lambda r) = C(\lambda) P(\geq r)\;. \tag{2}$$

Assuming that $P(\geq r)$ is a differentiable function, and differentiate Eq. (2) by λ, and then let $\lambda = 1$.

$$rP'(\geq r) = C'(1) P(\geq r) \tag{3}$$

As $C'(1)$ is a constant this differential equation is readily integrated as

$$P(\geq r) = c_0 r^{C'(1)}\,. \tag{4}$$

$P(\geq r)$ is a cumulative distribution and it is a non-increasing function in general, the exponent $C'(1)$ can be replaced by $-\alpha$ where α is a positive constant. Namely, from the scale-invariance with the assumption of differentiability we have the following power law:

$$P(\geq r) = c_0 r^{-\alpha} \,. \tag{5}$$

The reversed logic also holds, namely for any power law distribution there is a fractal configuration or a scale-invariant state.

In the case of real impact fracture, the size distribution of pieces is experimentally obtained by repeating sieves of various sizes, and it is empirically well-known that a fractured piece's diameter follows a power law with the exponent about $\alpha = 2$ independent of the details about the material or the way of impact [14]. This law is one of the most stubborn physical laws in nature as it is known to hold from 10^{-6} m to 10^5 m, from glass pieces around us to asteroids. From theoretical viewpoint this phenomenon is known to be described by a scale-free dynamics of crack propagation and the universal properties of the exponent value are well understood [19].

Usually fractal is considered geometric concept introducing the quantity fractal dimension or the concept of self-similarity. However, in economics there are very few geometric objects, so, the concept of fractals in economics are mostly used in the sense of power law distributions.

It should be noted that any geometrical fractal object accompanies a power law distribution even a deterministic fractal such as Sierpinski gasket. Figure 2 shows Sierpinski gasket which is usually characterized by the fractal dimension D given by

$$D = \frac{\log 3}{\log 2} \,. \tag{6}$$

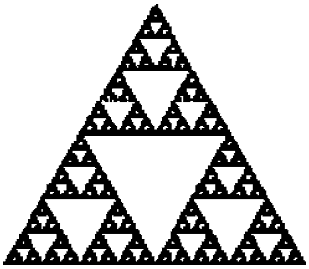
Paying attention to the distribution of length r of white triangles in this figure, it is easy to show that the probability that a randomly chosen white triangle's side is larger than r, $P(\geq r)$, follows the power law,

$$P(\geq r) \propto r^{-\alpha} \,, \quad \alpha = D = \frac{\log 3}{\log 2} \,. \tag{7}$$

Here, the power law exponent of distribution equals to the fractal dimension; however, such coincidence occurs only when the considering distribution is for a length distribution. For example, in Sierpinski gasket the area s of white triangles follow the power law,

$$P(\geq s) \propto s^{-\alpha} \,, \quad \alpha = \frac{\log 3}{\log 4} \,. \tag{8}$$

The fractal dimension is applicable only for geometric fractals, however, power law distributions are applicable for any fractal phenomena including shapeless quantities. In such cases the power law exponent is the most important quantity for quantitative characterization of fractals.



Fractals and Economics, Figure 2
Sierpinski gasket

According to Mandelbrot's own review on his life the concept of fractal was inspired when he was studying economics data [26]. At that time he found two basic properties in the time series data of daily prices of New York cotton market [24]:

(A) Geometrical similarity between large scale chart and an expanded chart.
(B) Power law distribution of price changes in a unit time interval, which is independent of the time scale of the unit.

He thought such scale invariance in both shape and distribution is a quite general property, not only in price charts but also in nature at large. His inspiration was correct and the concept of fractals spread over physics first and then over almost all fields of science. In the history of science it is a rare event that a concept originally born in economics has been spread widely to all area of sciences.

Basic mathematical properties of cumulative distribution can be summarized as follows (here we consider distribution of non-negative quantity for simplicity):

1. $P(\geq 0) = 1 \,, P(\geq \infty) = 0$.
2. $P(\geq r)$ is a non-increasing function of r.
3. The probability density is given as $p(r) \equiv -\frac{\mathrm{d}}{\mathrm{d}r} P(\geq r)$.
   As for power law distributions there are three peculiar characteristics:
4. Difficulty in normalization. Assuming that $P(\geq r) = c_0 r^{-\alpha}$ for all in the range $0 \leq r < \infty$, then the normalization factor c_0 must be 0 considering the limit of $r \to 0$. To avoid this difficulty it is generally assumed that the power law does not hold in the vicinity of $r = 0$. In the case of observing distribution from real data there are naturally lower and upper bounds, so this difficulty should be necessary only for theoretical treatment.

5. Divergence of moments. As for moments defined by $\langle r^n \rangle \equiv \int_0^\infty r^n p(r) \mathrm{d}r$, $\langle r^n \rangle = \infty$ for $n \geq \alpha$. In the special case of $2 \geq \alpha > 0$ the basic statistical quantity, the variance, diverges, $\sigma^2 \equiv \langle r^2 \rangle - \langle r \rangle^2 = \infty$. In the case of $1 \geq \alpha > 0$ even the average can not be defined as $\langle r \rangle = \infty$.
6. Stationary or non-stationary? In view of the data analysis, the above characteristics of diverging moments is likely to cause a wrong conclusion that the phenomenon is non-stationary by observing its averaged value. For example, assume that we observe k samples $\{r_1, r_2, \ldots, r_k\}$ independently from the power law distribution with the exponent, $1 \geq \alpha > 0$. Then, the sample average, $\langle r \rangle_k \equiv \frac{1}{k}\{r_1 + r_2 + \cdots + r_k\}$, is shown to diverge as, $\langle r \rangle_k \propto k^{1/\alpha}$. Such tendency of monotonic increase of averaged quantity might be regarded as a result of non-stationarity, however, this is simply a general property of a power law distribution. The best way to avoid such confusion is to observe the distribution directly from the data.

Other than the power law distribution there is another important statistical quantity in the study of fractals, that is, the autocorrelation. For given time series, $\{x(t)\}$, the autocorrelation is defined as,

$$C(T) \equiv \frac{\langle x(t+T)x(t) \rangle - \langle x(t) \rangle^2}{\langle x(t)^2 \rangle - \langle x(t) \rangle^2}\,, \tag{9}$$

where $\langle \cdots \rangle$ denotes an average over realizations. The autocorrelation can be defined only for stationary time series with finite variance, in which any statistical quantities do not depend on the location of the origin of time axis.

For any case, the autocorrelation satisfies the following basic properties,

1. $C(0) = 1$ and $C(\infty) = 0$
2. $|C(T)| \leq 1$ for any $T \geq 0$.
3. The Wiener–Khinchin theorem holds, $C(T) = \int_0^\infty S(f) \cos 2\pi f \mathrm{d}f$, where $S(f)$ is the power spectrum defined by $S(f) \equiv \langle \widehat{x}(f)\widehat{x}(-f) \rangle$, with the Fourier transform, $\widehat{x}(f) \equiv \int x(t) \mathrm{e}^{2\pi i f t} \mathrm{d}t$.

In the case that the autocorrelation function is characterized by a power law, $C(T) \propto T^{-\beta}$, $\beta > 0$, then the time series $\{x(t)\}$ is said to have a fractal property, in the sense that the autocorrelation function is scale-independent for any scale-factor, $\lambda > 0$, $C(\lambda T) \propto C(T)$. In the case $1 > \beta > 0$ the corresponding power spectrum is given as $S(f) \propto f^{-1+\beta}$.

The power spectrum can be applied to any time series including non-stationary situations. A simple way of telling non-stationary situation is to check the power law exponent of $S(f) \propto f^{-1+\beta}$ in the vicinity of $f = 0$, for $0 > \beta$ the time series is non-stationary.

Three basic examples of fractal time series are the followings:

1. White noise. In the case that $\{x(t)\}$ is a stationary independent noise, the autocorrelation is given by the Kronecker's function, $C(T) = \delta_T$, where
$$\delta_T = \begin{cases} 1\,, & T = 0 \\ 0\,, & T \neq 0\,. \end{cases}$$
The corresponding power spectrum is $S(f) \propto f^0$. This case is called white noise from an analogy that superposition of all frequency lights with the same amplitude make a colorless white light. White noise is a plausible model of random phenomena in general including economic activities.
2. Random walk. This is defined by summation of a white noise, $X(t) = X(0) + \sum_{s=0}^{t} x(s)$, and the power spectrum is given by $S(f) \propto f^{-2}$. In this case the autocorrelation function can not be defined because the data is non-stationary. Random walks are quite generic models widely used from Brownian motions of colloid to market prices. The graph of a random walk has a fractal property such that an expansion of any part of the graph looks similar to the whole graph.
3. The $1/f$ noise. The boundary of stationary and non-stationary states is given by the so-called $1/f$ noise, $S(f) \propto f^{-1}$. This type of power spectrum is also widely observed in various fields of sciences from electrical circuit noise [16] to information traffics in the Internet [53]. The graph of this $1/f$ noise also has the fractal property.

Examples in Economics

In this chapter fractals observed in real economic activities are reviewed. Mathematical models derived from these empirical findings will be summarized in the next chapter.

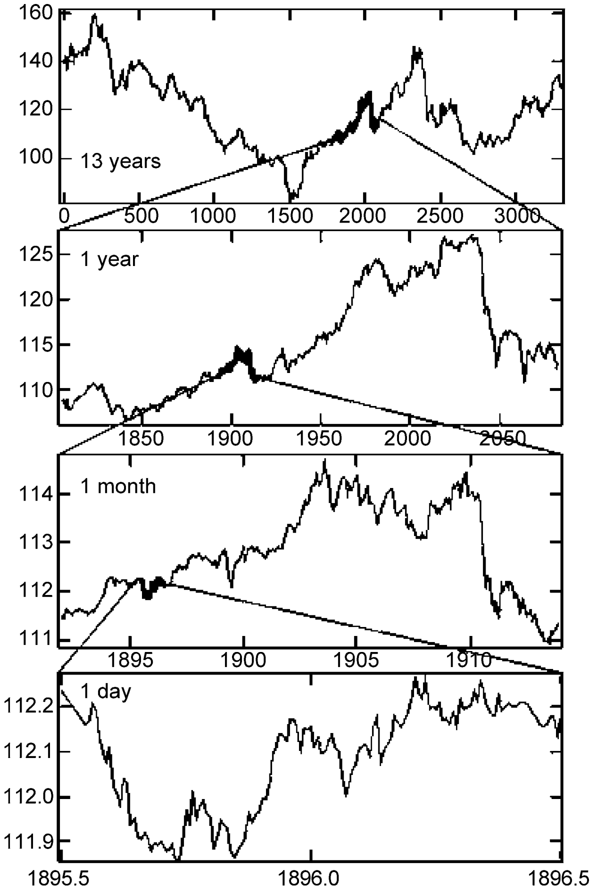
As mentioned in the previous chapter the very first example of a fractal was the price fluctuation of the New York cotton market analyzed by Mandelbrot with the daily data for a period of more than a hundred years [24]. This research attracted much attention at that time, however, there was no other good market data available for scientific analysis, and no intensive follow-up research was done until the 1990s. Instead of earnest scientific data analysis artificial mathematical models of market prices based on random walk theory became popular by the name of Financial Technology during the years 1960–1980.

Fractal properties of market prices are confirmed with huge amount of high resolution market data since the 1990s [26,43,44]. This is due to informationization of financial markets in which transaction orders are processed by computers and detail information is recorded automatically, while until the 1980s many people gathered at a market and prices are determined by shouting and screaming which could not be recorded. Now there are more than 100 financial market providers in the world and the number of transacted items exceeds one million. Namely, millions of prices in financial markets are changing with time scale in seconds, and you can access any market price at real time if you have a financial provider's terminal on your desk via the Internet.

Among these millions of items one of the most representative financial markets is the US Dollar-Japanese Yen (USD-JPY) market. In this market Dollar and Yen are exchanged among dealers of major international banks. Unlike the case of stock markets there is no physical trading place, but major international banks are linked by computer networks and orders are emitted from each dealer's terminal and transactions are done at an electronic broking system. Such a broking system and the computer networks are provided by financial provider companies like Reuters.

The foreign exchange markets are open 24 hours and deals are done whenever buy- and sell-orders meet. The minimum unit of a deal is one million USD (called a bar), and about three million bars are traded everyday in the whole foreign exchange markets in which more than 100 kinds of currencies are exchanged continuously. The total amount of money flow is about 100 times bigger than the total amount of daily world trade, so it is believed that most of deals are done not for the real world's needs, but they are based on speculative strategy or risk hedge, that is, to get profit by buying at a low price and selling at a high price, or to avoid financial loss by selling decreasing currency.

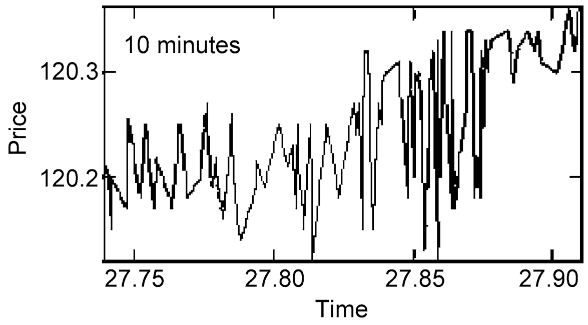
In Fig. 3 the price of one US Dollar paid by Japanese Yen in the foreign exchange markets is shown for 13 years [30]. The total number of data points is about 20 million, that is, about 10 thousand per day or the averaged transaction interval is seven seconds. A magnified part of the top figure for one year is shown in the second figure. The third figure is the enlargement of one month in the second figure. The bottom figure is again a part of the third figure, here the width is one day. It seems that at least the top three figures look quite similar. This is one of the fractal properties of market price (A) introduced in the previous chapter. This geometrical fractal property can be found in any market, so that this is a very universal market property.



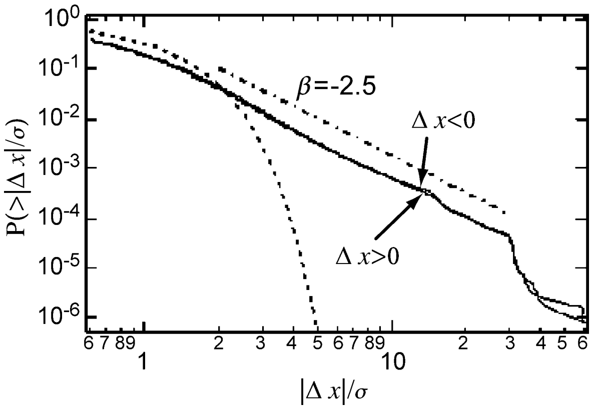
Fractals and Economics, Figure 3
Dollar-Yen rate for 13 years (*Top*). *Dark areas* are enlarged in the following figure [30]

However, it should be noted that this geometrical fractal property breaks down for very short time scale as typically shown in Fig. 4. In this figure the abscissa is 10 minutes range and we can observe each transaction separately. Obviously the price up down is more zigzag and more discrete than the large scale continuous market fluctuations shown in Fig. 3. In the case of USD-JPY market the time scale that this breakdown of scale invariance occurs typically at time scale of several hours.

The distribution of rate change in a unit time (one minute) is shown in Fig. 5. Here, there are two plots of cumulative distributions, $P(> \Delta x)$ for positive rate changes and $P(> |\Delta x|)$ for negative rate changes, which are almost identical meaning that the up-down symmetry of rate changes is nearly perfect. In this log–log plot the estimated power law distribution's exponent is 2.5. In the



Fractals and Economics, Figure 4
Market price changes in 10 minutes

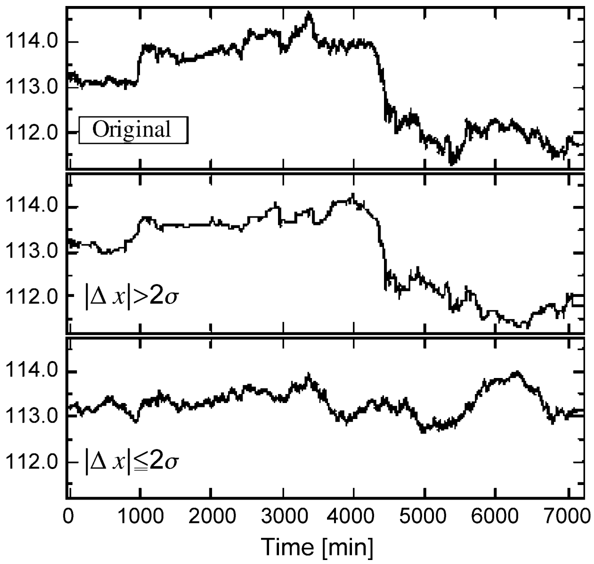


Fractals and Economics, Figure 5
Log–log plot of cumulative distribution of rate change [30]

original finding of Mandelbrot, (B) in the previous chapter, the reported exponent value is about 1.7 for cotton prices. In the case of stock markets power laws are confirmed universally for all items, however, the power exponents are not universal, taking value from near one to near five, typically around three [15]. Also the exponent values change in time year by year.

In order to demonstrate the importance of large fluctuations, Fig. 6 shows a comparison of three market prices. The top figure is the original rate changes for a week. The middle figure is produced from the same data, but it is consisted of rate changes of which absolute values are larger than 2σ, that is, about 5 % of all the data. In the bottom curve such large rate changes are omitted and the residue of 95 % of small changes makes the fluctuations. As known from these figures the middle figure is much closer to the original market price changes. Namely, the contribution from the power law tails of price change distribution is very large for macro-scale market prices.

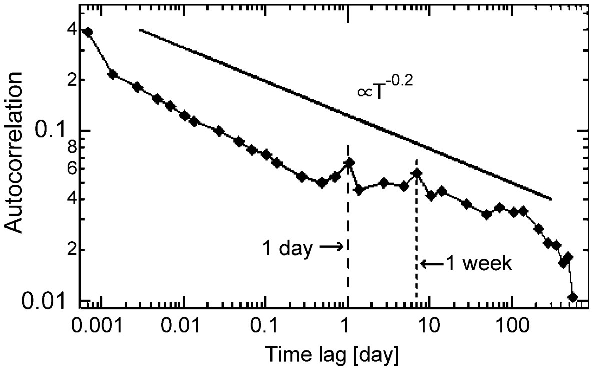
Power law distribution of market price changes is also a quite general property which can be confirmed for any market. Up-down symmetry also holds universally in short time scale in general, however, for larger unit time the distribution of price changes gradually deforms and for very large unit time the distribution becomes closer to a Gaussian distribution. It should be noted that in special cases of market crashes or bubbles or hyper-inflations the up-down symmetry breaks down and the power law distribution is also likely to be deformed.



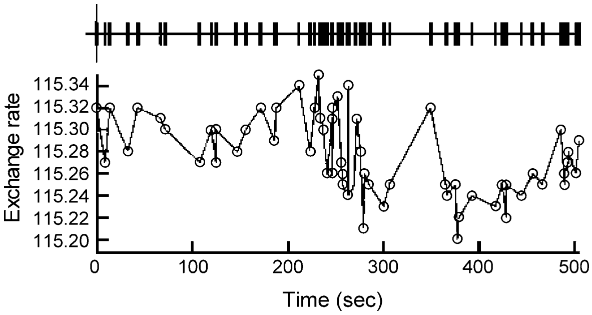
Fractals and Economics, Figure 6
USD-JPY exchange rate for a week (*top*) Rate changes smaller than 2σ are neglected (*middle*) Rate changes larger than 2σ are neglected (*bottom*)

The autocorrelation of the time sequence of price changes generally decays quickly to zero, sometimes accompanied by a negative correlation in a very short time. This result implies that the market price changes are apparently approximated by white noise, and market prices are known to follow nearly a random walk as a result. However, market price is not a simple random walk. In Fig. 7 the autocorrelation of volatility, which is defined by the square of price change, is shown in log–log scale. In the case of a simple random walk this autocorrelation should also decay quickly. The actual volatility autocorrelation nearly satisfies a power law implying that the volatility time series has a fractal clustering property. (See also Fig. 31 representing an example of price change clustering.)

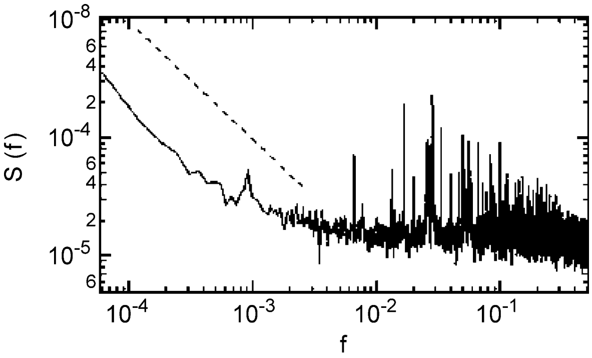
Another fractal nature of markets can be found in the intervals of transactions. As shown in Fig. 8 the transaction intervals fluctuate a lot in very short time scale. It is known that the intervals make clusters, namely, shorter in-



Fractals and Economics, Figure 7
Autocorrelation of volatility [30]



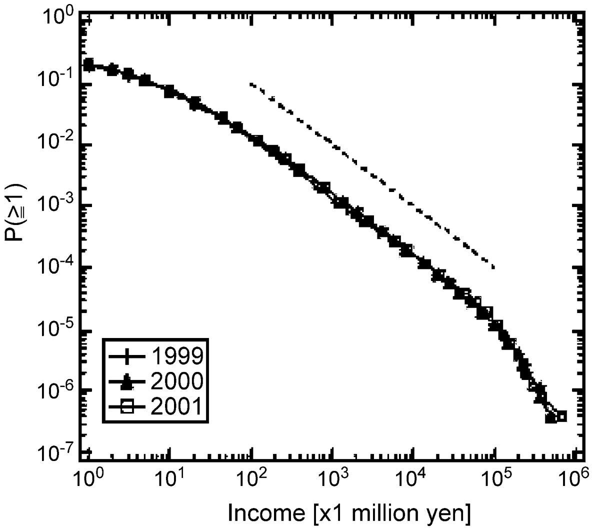
Fractals and Economics, Figure 8
Clustering of transaction intervals



Fractals and Economics, Figure 9
Power spectrum of transaction intervals [50]

tervals tend to gather. To characterize such clustering effect we can make a time sequence consisted of 0 and 1, where 0 denotes no deal was done at that time, and 1 denotes a deal was done. The corresponding power spectrum follows a $1/f$ power spectrum as shown in Fig. 9 [50].

Fractal properties are found not only in financial markets. Company's income distribution is known to follow



Fractals and Economics, Figure 10
Income distribution of companies in Japan

also a power law [35]. A company's income is roughly given by subtraction of incoming money flow minus outgoing money flow, which can take both positive and negative values. There are about six million companies in Japan and Fig. 10 shows the cumulative distribution of annual income of these companies. Clearly we have a power law distribution of income I with the exponent very close to -1 in the middle size range, so-called the Zipf's law,

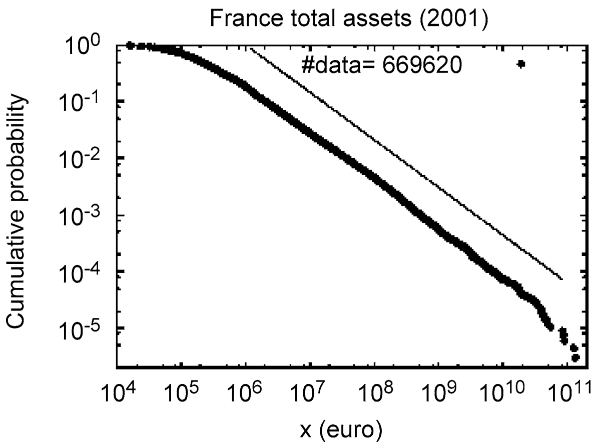
$$P(> I) \propto I^{-\beta}\,, \quad \beta = 1\,. \tag{10}$$

Although in each year every company's income fluctuates, and some percentage of companies disappear or are newly born, this power law is known to hold for more than 30 years. Similar power laws are confirmed in various countries, the case of France is plotted in Fig. 11 [13].

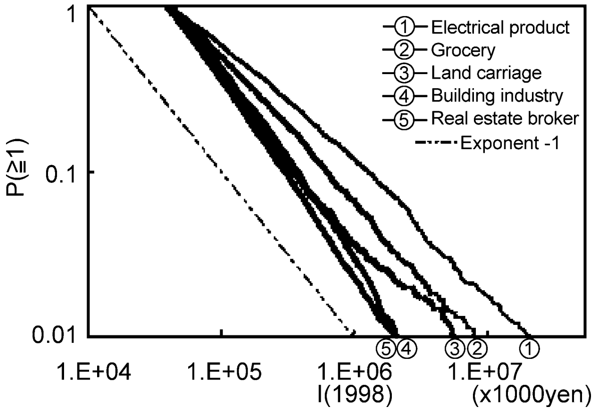
Observing more details by categorizing the companies, it is found that the income distribution in each job category follows nearly a power law with the exponent depending on the job category as shown in Fig. 12 [29]. The implication of this phenomenon will be discussed in Sect. "Income Distribution Models".

A company's size can also be viewed by the amount of whole sale or the number of employee. In Figs. 13 and 14 distributions of these quantities are plotted [34]. In both cases clear power laws are confirmed. The size distribution of debts of bankrupted companies is also known to follow a power law as shown Fig. 15 [12].

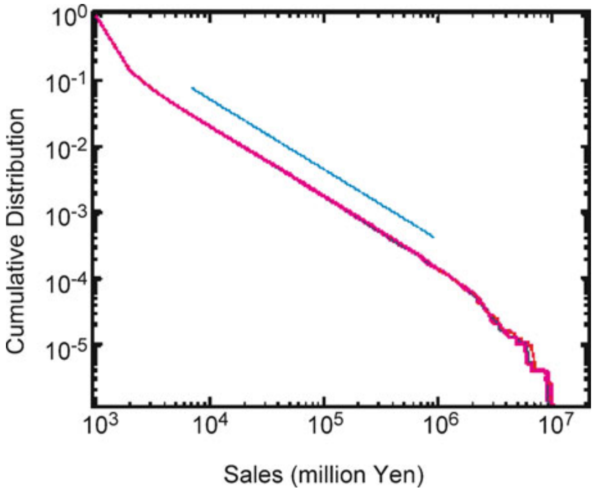
A power law distribution can also be found in personal income. Figure 16 shows the personal income distribution in Japan in a log–log plot [1]. The distribution



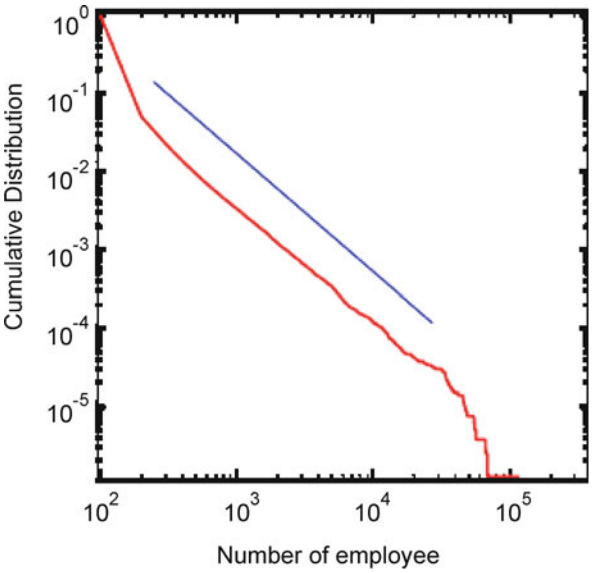
Fractals and Economics, Figure 11
Income distribution of companies in France [13]



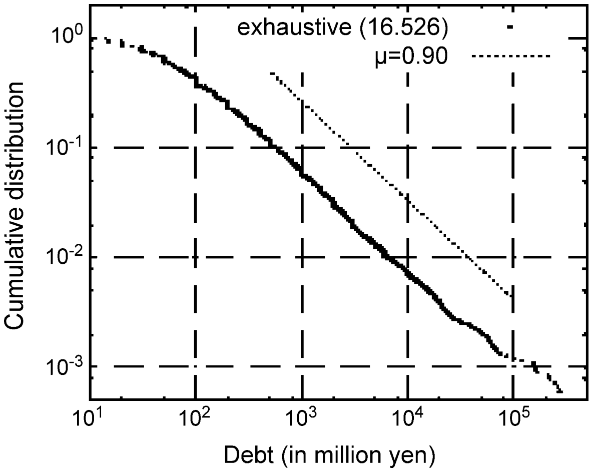
Fractals and Economics, Figure 12
Income distribution of companies in each category [29]



Fractals and Economics, Figure 13
The distribution of whole sales [34]



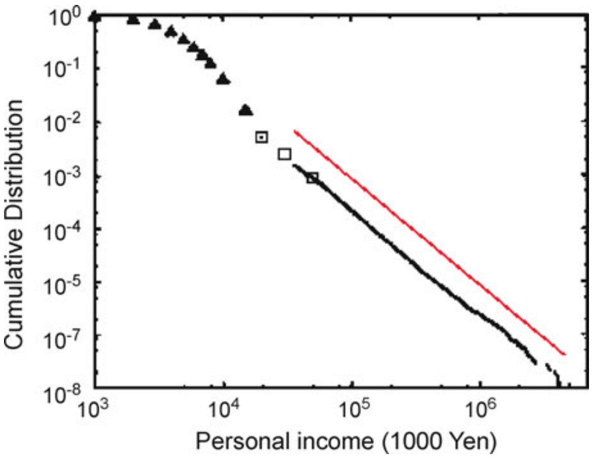
Fractals and Economics, Figure 14
The distribution of employee numbers [34]



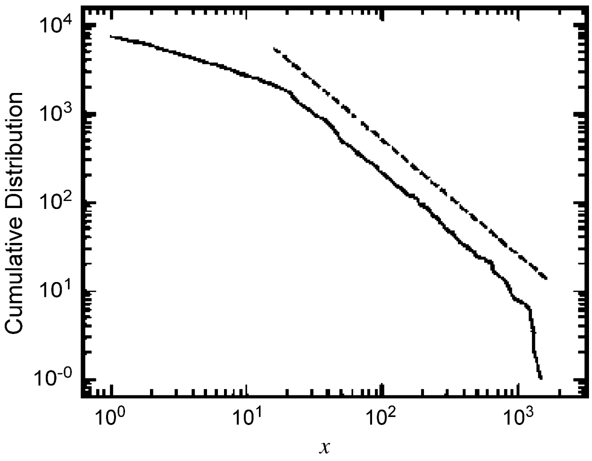
Fractals and Economics, Figure 15
The size distribution of debts of bankrupted companies [12]

is clearly separated into two parts. The majority of people's incomes are well approximated by a log-normal distribution (the left top part of the graph), and the top few percent of people's income distribution is nicely characterized by a power law (the linear line in the left part of the graph). The majority of people are getting salaries from companies. This type of composite of two distributions is well-known from the pioneering study by Pareto about 100 years ago and it holds in various countries [8,22].

A typical value of the power exponent is about two, significantly larger than the income distribution of com-



Fractals and Economics, Figure 16
Personal income distribution in Japan [1]



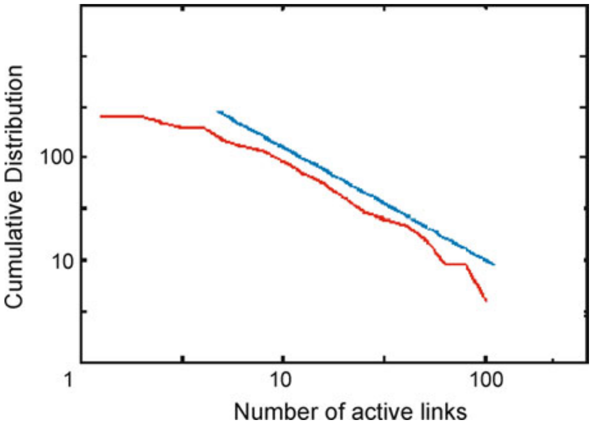
Fractals and Economics, Figure 17
The distribution of the amount of transferred money [21]

panies. However, the exponent of the power law seems to be not universal and the value changes county by country or year by year. There is a tendency that the exponent is smaller, meaning more rich people, when the economy is improving [40].

Another fractal in economics can be found in a network of economic agents such as banks' money transfer network. As a daily activity banks transfer money to other banks for various reasons. In Japan all of these interbank money transfers are done via a special computer network provided by the Bank of Japan. Detailed data of actual money transfer among banks are recorded and analyzed for the basic study.

The total amount of money flow among banks in a day is about 30 × 10^{12} yen with the number of transactions about 10 000. Figure 17 shows the distribution of the amount of money at a transaction. The range is not wide enough but we can find a power law with an exponent about 1.3 [20].

The number of banks is about 600, so the daily transaction number is only a few percent of the theoretically possible combinations. It is confirmed that there are many pairs of banks which never transact directly. We can define active links between banks for pairs with the averaged number of transaction larger than one per day. By this criterion the number of links becomes about 2000, that is, about 0.5 percent of all possible link numbers. Compared with the complete network, the actual network topologies are much more sparse.

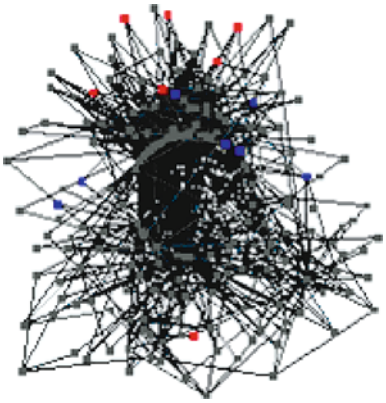


Fractals and Economics, Figure 18
The number distribution of active links per site [20]

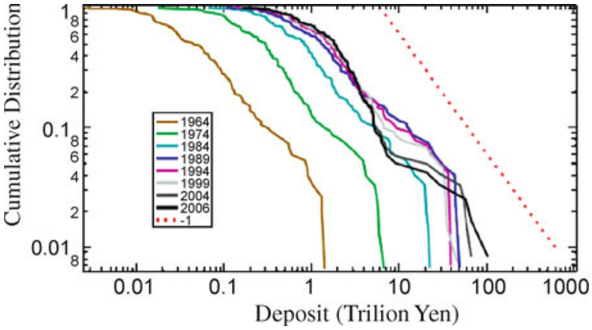
In Fig. 18 the number distribution of active links per site are plotted in log–log plot [21]. As is known from this graph, there is an intermediate range in which the link number distribution follows a power law. In the terminology of recent complex network study, this property is called the scale-free network [5]. The scale-free network structure among these intermediate banks is shown in Fig. 19.

There are about 10 banks with large link numbers which deviate from the power law, also small link number banks with link number less than four are out of the power law. Such small banks are known to make a satellite structure that many banks linked to one large link number banks. It is yet to clarify why intermediate banks make fractal network, and also to clarify the role of large banks and small banks which are out of the fractal configuration.

In relation with the banks, there are fractal properties other than cash flow and the transaction network. The distribution of the whole amount of deposit of Japanese bank



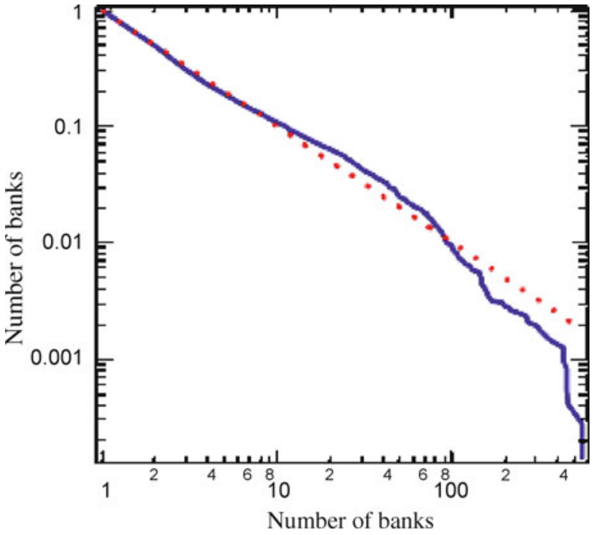
Fractals and Economics, Figure 19
Scale-free network of intermediate banks [20]



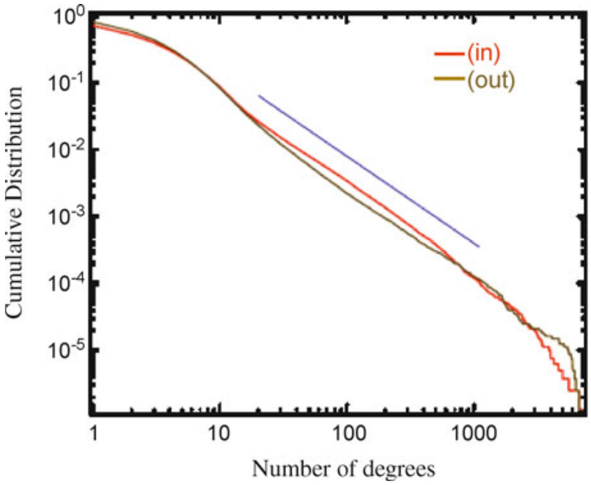
Fractals and Economics, Figure 20
Distribution of total deposit for Japanese banks [57] Power law breaks down from 1999

is approximated by a power law as shown in Fig. 20 [57]. In recent years large banks merged making a few mega banks and the distribution is a little deformed. Historically there were more than 6000 banks in Japan, however, now we have about 600 as mentioned. It is very rare that a bank disappears, instead banks are merged or absorbed. The number distribution of banks which are historically behind a present bank is plotted in Fig. 21, again a power law can be confirmed.

Other than the example of the bank network, network structures are very important generally in economics. In production process from materials, through various parts to final products the network structure is recently studied in view of complex network analysis [18]. Trade networks among companies can also be described by network terminology. Recently, network characterization quantities such as link numbers (Fig. 22), degrees of authority, and Pageranks are found to follow power laws from real trade data for nearly a million of companies in Japan [34].



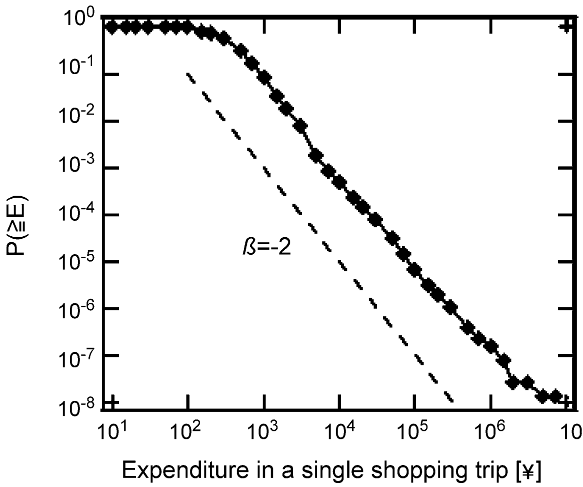
Fractals and Economics, Figure 21
Distribution of bank numbers historically behind a present bank [57]



Fractals and Economics, Figure 22
Distribution of in-degrees and out-degrees in Japanese company network [34]

Still more power laws in economics can be found in sales data. A recent study on the distribution of expenditure at convenience stores in one shopping trip shows a clear power law distribution with the exponent close to two as shown in Fig. 23 [33]. Also, book sales, movie hits, news paper sales are known to be approximated by power laws [39].

Viewing all these data in economics, we may say that fractals are everywhere in economics. In order to under-



Fractals and Economics, Figure 23
Distribution of expenditure in one shopping trip [33]

stand why fractals appear so frequently, we firstly need to make simple toy models of fractals which can be analyzed completely, and then, based on such basic models we can make more realistic models which can be directly comparable with real data. At that level of study we will be able to predict or control the complex real world economy.

Basic Models of Power Laws

In this chapter we introduce general mathematical and physical models which produce power law distributions. By solving these simple and basic cases we can deepen our understanding of the underlying mechanism of fractals or power law distributions in economics.

Transformation of Basic Distributions

A power law distribution can be easily produced by variable transformation from basic distributions.

1. Let x be a stochastic variable following a uniform distribution in the range $(0, 1]$, then, $y \equiv x^{-1/\alpha}$ satisfies a power law, $P(> y) = y^{-\alpha}$ for $y \geq 1$. This is a useful transformation in case of numerical simulation using random variable following power laws.
2. Let x be a stochastic variable following an exponential distribution, $P(> x) = \mathrm{e}^{-x}$, for positive x, then, $y \equiv \mathrm{e}^{x/\alpha}$ satisfies a power law, $P(> y) \propto y^{-\alpha}$. As exponential distributions occur frequently in random process such as the Poisson process, or energy distribution in thermal equilibrium, this simple exponential variable transformation can make it a power law.

Superposition of Basic Distributions

A power law distribution can also be easily produced by superposition of basic distributions.

Let x be a Gaussian distribution with the probability density given by

$$p_R(x) = \frac{\sqrt{R}}{\sqrt{2\pi}} \mathrm{e}^{-\frac{R}{2}x^2} , \tag{11}$$

and R be a χ^2 distribution with degrees of freedom α,

$$w(R) = \frac{\left(\frac{1}{2}\right)^{\alpha/2}}{\Gamma\left(\frac{\alpha}{2}\right)} R^{\frac{\alpha}{2}-1} \mathrm{e}^{-\frac{R}{2}} . \tag{12}$$

Then, the superposition of Gaussian distribution, Eq. (11), with the weight given by Eq. (12) becomes the T-distribution having power law tails:

$$\begin{aligned} p(x) &= \int_0^\infty W(R) p_R(x) \mathrm{d}R \\ &= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{(1+x^2)^{\frac{\alpha+1}{2}}} \propto |x|^{-\alpha-1} , \end{aligned} \tag{13}$$

which is $P(> |x|) \propto |x|^{-\alpha}$ in cumulative distribution. In the special case that R, the inverse of variance of the normal distribution, distributes exponentially, the value of α is 2. Similar super-position can be considered for any basic distributions and power law distributions can be produced by such superposition.

Stable Distributions

Assume that stochastic variables, $x_1, x_2, \ldots, x_n$, are independent and follow the same distribution, $p(x)$, then consider the following normalized summation;

$$X_n \equiv \frac{x_1 + x_2 + \cdots + x_n - \mu_n}{n^{1/\alpha}} . \tag{14}$$

If there exists $\alpha > 0$ and μ_n, such that the distribution of X_n is identical to $p(x)$, then, the distribution belongs to one of the Levy stable distributions [10]. The parameter α is called the characteristic exponent which takes a value in the range (0, 2]. The stable distribution is characterized by four continuous parameters, the characteristic exponent, an asymmetry parameter which takes a value in $[-1, 1]$, the scale factor which takes a positive value and the location parameter which takes any real number. Here, we introduce just a simple case of symmetric distribution around the origin with the unit scale factor. The probability density is then given as

$$p(x;\alpha)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-i\rho x}\mathrm{e}^{-|\rho|^{\alpha}}\mathrm{d}\rho\,. \tag{15}$$

For large $|x|$ the cumulative distribution follows the power law, $P(>x;\alpha)\propto|x|^{-\alpha}$ except the case of $\alpha=2$. The stable distribution with $\alpha=2$ is the Gaussian distribution.

The most important property of the stable distribution is the generalized central limit theorem: If the distribution of sum of any independent identically distributed random variables like X_n in Eq. (14) converges in the limit of $n\to\infty$ for some value of α, then the limit distribution is a stable distribution with the characteristic exponent α. For any distribution with finite variance, the ordinary central limit theory holds, that is, the special case of $\alpha=2$. For any infinite variance distribution the limit distribution is $\alpha\neq2$ with a power law tail. Namely, a power law realizes simply by summing up infinitely many stochastic variables with diverging variance.

Entropy Approaches

Let x_0 be a positive constant and consider a probability density $p(x)$ defined in the interval $[x_0,\infty)$, the entropy of this distribution is given by

$$S\equiv-\int_{x_0}^{\infty}p(x)\log p(x)\mathrm{d}x\,. \tag{16}$$

Here, we find a distribution that maximizes the entropy with a constraint such that the expectation of logarithm of x is a constant, $\langle\log x\rangle=M$. Then, applying the variational principle to the following function,

$$L\equiv-\int_{x_0}^{\infty}p(x)\log p(x)\mathrm{d}x-\lambda_1\left(\int_{x_0}^{\infty}p(x)\mathrm{d}x-1\right)+\lambda_2\left(\int_{x_0}^{\infty}p(x)\log x\mathrm{d}x-M\right) \tag{17}$$

the power law is obtained,

$$P(\geq x)=\left(\frac{x}{x_0}\right)^{-\frac{1}{M-\log x_0}}\,. \tag{18}$$

In other words, a power law distribution maximizes the entropy in the situation where products are conserved. To be more precise, consider two time dependent random variables interacting each other satisfying the relation, $x_1(t)\cdot x_2(t)=x_1(t')\cdot x_2(t')$, then the equilibrium distribution follows a power law.

Another entropy approach to the power laws is to generalize the entropy by the following form [56],

$$S_q\equiv\frac{1-\int_{x_0}^{\infty}p(x)^q\mathrm{d}x}{q-1}\,, \tag{19}$$

where q is a real number. This function is called the q-entropy and the ordinary entropy, Eq. (15), recovers in the limit of $q\to1$. Maximizing the q-entropy keeping the variance constant, so-called a q-Gaussian distribution is obtained, which has the same functional form with the T-distribution, Eq. (12), with the exponent α given by

$$\alpha=\frac{q-3}{1-q}\,. \tag{20}$$

This generalized entropy formulation is often applied to nonlinear systems having long correlations, in which power law distributions play the central role.

Random Multiplicative Process

Stochastic time evolution described by the following formulation is called the multiplicative process,

$$x(t+1)=b(t)x(t)+f(t)\,, \tag{21}$$

where $b(t)$ and $f(t)$ are both independent random variables [17]. In the case that $b(t)$ is a constant, the distribution of $x(t)$ depends on the distribution of $f(t)$, for example, if $f(t)$ follows a Gaussian distribution, then the distribution of $x(t)$ is also a Gaussian. However, in the case that $b(t)$ fluctuates randomly, the resulting distribution of $x(t)$ is known to follows a power law independent of $f(t)$,

$$P(>x)\propto|x|^{-\alpha}\,, \tag{22}$$

where the exponent α is determined by solving the following equation [48],

$$\langle|b(t)|^{\alpha}\rangle=1\,. \tag{23}$$

This steady distribution exists when $\langle\log|b(t)|\rangle<0$ and $f(t)$ is not identically 0. As a special case that $b(t)=0$ with a finite probability, then a steady state exists. It is proved rigorously that there exists only one steady state, and starting from any initial distribution the system converges to the power law steady state.

In the case $\langle\log|b(t)|\rangle\geq0$ there is no statistically steady state, intuitively the value of $|b(t)|$ is so large that $x(t)$ is likely to diverge. Also in the case $f(t)$ is identically 0 there is no steady state as known from Eq. (21) that

$\log|x(t)|$ follows a simple random walk with random noise term, $\log|b(t)|$.

The reason why this random multiplicative process produces a power law can be understood easily by considering a special case that $b(t) = b > 1$ with probability 0.5 and $b(t) = 0$ otherwise, with a constant value of $f(t) = 1$. In such a situation the value of $x(t)$ is $1 + b + b^2 + \cdots + b^K$ with probability $(0.5)^K$. From this we can directly evaluate the distribution of $x(t)$,

$$P\left(\geq \frac{b^{K+1}-1}{b-1}\right) = 2^{-K+1} \quad \text{i. e.} \qquad (24)$$
$$P(\geq x) = 4(1+(b-1)x)^{-\alpha}\,, \quad \alpha = \frac{\log 2}{\log b}\,.$$

As is known from this discussion, the mechanism of this power law is deeply related to the above mentioned transformation of exponential distribution in Sect. "Transformation of Basic Distributions".

The power law distribution of a random multiplicative process can also be confirmed experimentally by an electrical circuit in which resistivity fluctuates randomly [38]. In an ordinary electrical circuit the voltage fluctuations in thermal equilibrium is nearly Gaussian, however, for a circuit with random resistivity a power law distribution holds.

Aggregation with Injection

Assume the situation that many particles are moving randomly and when two particles collide they coalesce making a particle with mass conserved. Without any injection of particles the system converges to the trivial state that only one particle remains. In the presence of continuous injection of small mass particles there exists a non-trivial statistically steady state in which mass distribution follows a power law [41]. Actually, the mass distribution of aerosol in the atmosphere is known to follow a power law in general [11].

The above system of aggregation with injection can be described by the following model. Let j be the discrete space, and $x_j(t)$ be the mass on site j at time t, then choose one site and let the particle move to another site and particles on the visited site merge, then add small mass particles to all sites, this process can be mathematically given as,

$$x_j(t+1) = \begin{cases} x_j(t) + x_k(t) + f_j(t)\,, & \text{prob} = 1/N \\ x_j(t) + f_j(t)\,, & \text{prob} = (N-2)/N \\ f_j(t)\,, & \text{prob} = 1/N \end{cases} \qquad (25)$$

where N is the total number of sites and $f_j(t)$ is the injected mass to the site j.

The characteristic function, $Z(\rho, t) \equiv \langle e^{-\rho x_j(t)} \rangle$, which is the Laplace transform of the probability density, satisfies the following equation by assuming uniformity,

$$Z(\rho, t+1) = \left\{ \frac{N-2}{N} Z(\rho,t)^2 + \frac{1}{N} Z(\rho,t) + \frac{1}{N} \right\} \langle e^{-\rho f_j(t)} \rangle\,. \qquad (26)$$

The steady state solution in the vicinity of $\rho = 0$ is obtained as

$$Z(\rho) = 1 - \sqrt{\langle f \rangle}\,\rho^{1/2} + \cdots\,. \qquad (27)$$

From this behavior the following power law steady distribution is obtained.

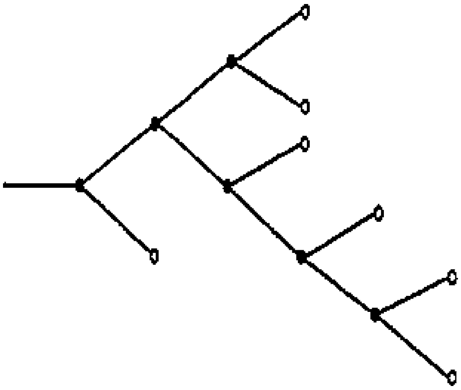
$$P(\geq x) \propto x^{-\alpha}\,, \quad \alpha = \frac{1}{2}\,. \qquad (28)$$

By introducing a collision coefficient depending on the size of particles power laws with various values of exponents realized in the steady state of such aggregation with injection system [46].

Critical Point of a Branching Process

Consider the situation that a branch grows and splits with probability q or stops growing with probability $1-q$ as shown in Fig. 24. What is the size distribution of the branch? This problem can be solved in the following way. Let $p(r)$ be the probability of finding a branch of size r, then the next relation holds.

$$p(r+1) = q \sum_{s=1}^{r-1} p(s)p(r-s)\,. \qquad (29)$$



Fractals and Economics, Figure 24
Branching process (from *left* to *right*)

Multiplying y^{r+1} and summing up by r from 0 to ∞, a closed equation of the generating function, $M(y)$, is obtained,

$$M(y)-1+q = q\cdot y\cdot M(y)^2\,, \quad M(y) \equiv \sum_{r=0}^{\infty} y^r p(r)\,. \quad (30)$$

Solving this quadratic equation and expanding in terms of y, we have the probability density,

$$p(r) \propto r^{-3/2} \mathrm{e}^{-Q(q)r}\,, \quad Q(q) \equiv \log 4q(1-q)\,. \quad (31)$$

For $q < 0.5$ the probability decays exponentially for large r, in this case all branches has a finite size. At $q = 0.5$ the branch size follows the power law, $P(\geq r) \propto r^{-1/2}$, and the average size of branch becomes infinity. For $q > 0.5$ there is a finite probability that a branch grows infinitely. The probability of having an infinite branch, $p(\infty) = 1-M(1)$, is given as,

$$p(\infty) = \frac{2q-1+\sqrt{1-4q(1-q)}}{2q}\,, \quad (32)$$

which grows monotonically from zero to one in the range $q = [0.5, 1]$. It should be noted that the power law distribution realizes at the critical point between the finite-size phase and the infinite-size phase [42].

Compared with the preceding model of aggregation with injection, Eq. (28), the mass distribution is the same as the branch size distribution at the critical point in Eq. (31). This coincidence is not an accident, but it is known that aggregation with injection automatically chooses the critical point parameter. Aggregation and branching are reversed process and the steady occurrence of aggregation implies that branching numbers keep a constant value on average and this requires the critical point condition. This type of critical behaviors is called the self-organized criticality and examples are found in various fields [4].

Finite Portion Transport

Here, a kind of mixture of aggregation and branching is considered. Assume that conserved quantities are distributed in N-sites. At each time step choose one site randomly, and transport a finite portion, $\theta x_j(t)$, to another randomly chosen site, where θ is a parameter in the range $[0, 1]$.

$$x_j(t+1) = (1-\theta)x_j(t)\,,$$
$$x_k(t+1) = x_k(t) + \theta x_j(t)\,. \quad (33)$$

It is known that for small positive θ the statistically steady distribution x is well approximated by a Gaussian like the case of thermal fluctuations. For θ close to 1 the fluctuation of x is very large and its distribution is close to a power law. In the limit θ goes to 1 and the distribution converges to Eq. (28), the aggregation with injection case. For intermediate values of θ the distribution accompanies a fat tail between Gaussian and a power law [49].

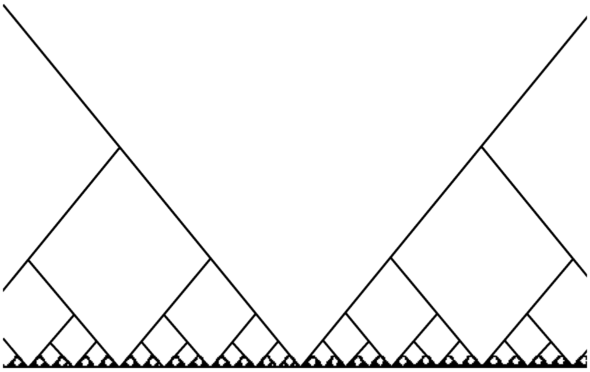
Fractal Tiling

A fractal tiling is introduced as the final basic model. Figure 25 shows an example of fractal tiling of a plane by squares. Like this case Euclidean space is covered by various sizes of simple shapes like squares, triangles, circles etc. The area size distribution of squares in Fig. 25 follows the power law,

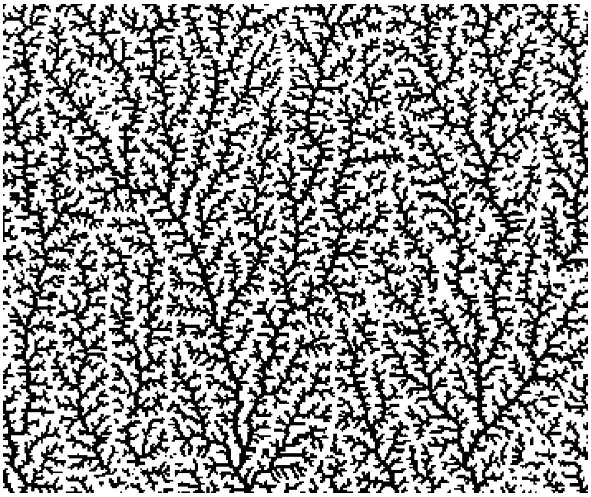
$$P(\geq x) \propto x^{-\alpha}\,, \quad \alpha = 1/2\,. \quad (34)$$

Generalizing this model in d-dimensional space, the distribution of d-dimensional volume x is characterized by a power law distribution with an exponent, $\alpha = (d-1)/d$, therefore, the Zipf's law which is the case of $\alpha = 1$ realizes in the limit of $d = \infty$. The fracture size distribution measured in mass introduced in the beginning of this article corresponds to the case of $d = 3$.

A classical example of fractal tiling is the Apollonian gasket, that is, a plane is covered totally by infinite number of circles which are tangent each other. For a given river pattern like Fig. 26 the basin area distribution follows a power law with exponent about $\alpha = 0.4$ [45]. Although these are very simple geometric models, simple models may sometimes help our intuitive understanding of fractal phenomena in economics.



Fractals and Economics, Figure 25
An example of fractal tiling



Fractals and Economics, Figure 26
Fractal tiling by river patterns [45]

Market Models

In this chapter market price models are reviewed in view of fractals. There are two approaches for construction of market models. One is modeling the time sequences directly by some stochastic model, and the other is modeling markets by agent models which are artificial markets in computer consisted of programmed dealers.

The first market price model was proposed by Bachelier in 1900 written as his Ph.D thesis [3], that is, five years before the model of Einstein's random walk model of colloid particles. His idea was forgotten for nearly 50 years. In 1950's Markowitz developed the portfolio theory based on a random walk model of market prices [28]. The theory of option prices by Black and Scholes was introduced in the 1970s, which is also based on random walk model of market prices, or to be more precise a logarithm of market prices in continuum description [7].

In 1982 Engle introduced a modification of the simple random walk model, the ARCH model, which is the abbreviation of auto-regressive conditional heteroscedasticity [9]. This model is formulated for market price difference as,

$$\Delta x(t) = \sigma(t) f(t) \,, \tag{35}$$

where $f(t)$ is a random variable following a Gaussian distribution with 0 mean and variance unity, the local variance $\sigma(t)$ is given as

$$\sigma(t)^2 = c_0 + \sum_{j=1}^{k} c_k (\Delta x(t-k))^2 \,, \tag{36}$$

with adjustable positive parameters, $\{c_0, c_1, \ldots, c_k\}$. By the effect of this modulation on variance, the distribution of price difference becomes superposition of Gaussian distribution with various values of variance, and the distribution becomes closer to a power law. Also, volatility clustering occurs automatically so that the volatility autocorrelation becomes longer.

There are many variants of ARCH models, such as GARCH and IGARCH, but all of them are based on purely probabilistic modeling, and the probability of prices going up and that of going down are identical.

Another type of market price model has been proposed from physics view point [53]. The model is called the PUCK model, an abbreviation of potentials of unbalanced complex kinetics, which assumes the existence of market's time-dependent potential force, $U_{\mathrm{M}}(x;t)$, and the time evolution of market price is given by the following set of equations;

$$x(t+1) - x(t) = -\frac{\mathrm{d}}{\mathrm{d}x} U_{\mathrm{M}}(x;t)\bigg|_{x=x(t)-x_{\mathrm{M}}(t)} + f(t) \,, \tag{37}$$

$$U_{\mathrm{M}}(x;t) \equiv \frac{b(t)}{M-1} \frac{x^2}{2} \,, \tag{38}$$

where M is the number of moving average needed to define the center of potential force,

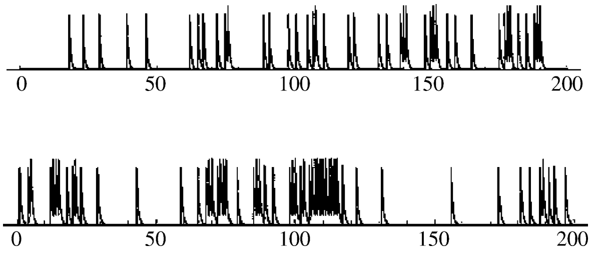
$$x_{\mathrm{M}}(t) \equiv \frac{1}{M} \sum_{k=0}^{M-1} x(t-k) \,. \tag{39}$$

In this model $f(t)$ is the external noise and $b(t)$ is the curvature of quadratic potential which changes with time. When $b(t) = 0$ the model is identical to the simple random walk model. When $b(t) > 0$ the market prices are attracted to the moving averaged price, $x_{\mathrm{M}}(t)$, the market is stable, and when $b(t) < 0$ prices are repelled from $x_{\mathrm{M}}(t)$ so that the price fluctuation is large and the market is unstable. For $b(t) < -2$ the price motion becomes an exponential function of time, which can describe singular behavior such as bubbles and crashes very nicely.

In the simplest case of $M = 2$ the time evolution equation becomes,

$$\Delta x(t+1) = -\frac{b(t)}{2} \Delta x(t) + f(t) \,. \tag{40}$$

As is known from this functional form in the case $b(t)$ fluctuates randomly, the distribution of price difference follows a power law as mentioned in the previous Sect. "Random Multiplicative Process", Random multiplicative process. Especially the PUCK model derives the ARCH model by introducing a random nonlinear potential function [54]. The value of $b(t)$ can be estimated from the



Fractals and Economics, Figure 27
Tick intervals of Poisson process (*top*) and the self-modulation process (*bottom*) [52]

data and most of known empirical statistical laws including fractal properties are fulfilled as a result [55].

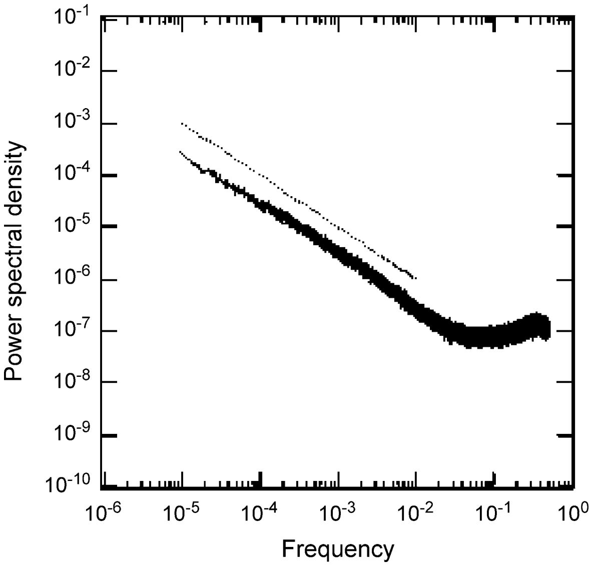
The peculiar difference of this model compared with financial technology models is that directional prediction is possible in some sense. Actually, from the data it is known that $b(t)$ changes slowly in time, and for non-zero $b(t)$ the autocorrelation is not zero implying that the up-down statistics in the near future is not symmetric. Moreover in the case of $b(t) < -2$ the price motion show an exponential dynamical growth hence predictable.

As introduced in Sect. "Examples in Economics" the tick interval fluctuations can be characterized by the $1/f$ power spectrum. This power law can be explained by a model called the self-modulation model [52]. Let Δt_j be the jth tick interval, and we assume that the tick interval can be approximated by the following random process,

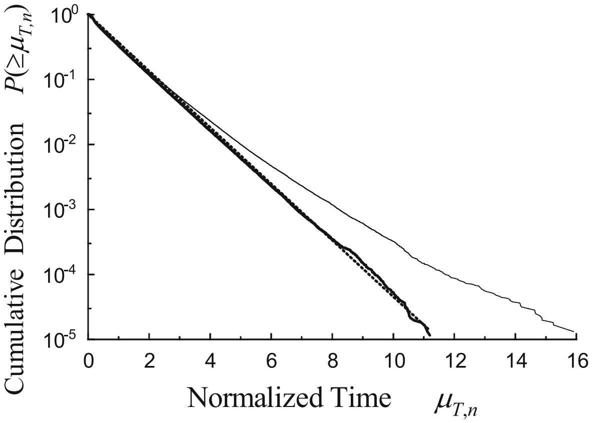
$$\Delta t_{j+1} = \mu_j \frac{1}{K} \sum_{k=0}^{K-1} \Delta t_{j-k} + g_j \,, \tag{41}$$

where μ_j is a positive random number following an exponential distribution with the mean value 1, and K is an integer which means the number of moving average, g_j is a positive random variable. Due to the moving average term in Eq. (41) the tick interval automatically make clusters as shown in Fig. 27, and the corresponding power spectrum is proved to be proportional to $1/f$ as typically represented in Fig. 28.

The market data of tick intervals are tested whether Eq. (41) really works or not. In Fig. 29 the cumulative probability of estimated value of μ_j from market data is plotted where the moving average size is determined by the physical time of 150 seconds and 400 seconds. As known from this figure, the distribution fits very nicely with the exponential distribution when the moving average size is 150 seconds. This result implies that dealers in the market are mostly paying attention to the latest transaction for about a few minutes only. And the dealers' clocks in their minds move quicker if the market becomes busier. By this self-modulation effect transactions in markets automatically make a fractal configuration.



Fractals and Economics, Figure 28
The power spectrum of the self-modulation process [52]



Fractals and Economics, Figure 29
The distribution of normalized time interval [50]

Next, we introduce a dealer model approach to the market [47]. In any financial market dealers' final goal is to gain profit from the market. To this end dealers try to buy at the lowest price and to sell at the highest price. Assume that there are N dealers at a market, and let the jth dealer's buying and selling prices in their mind $B_j(t)$ and $S_j(t)$. For each dealer the inequality, $B_j(t) < S_j(t)$, always holds. We pay attention to the maximum price of $\{B_j(t)\}$ called the best bid, and to the minimum price of $\{S_j(t)\}$ called the best ask. Transactions occur in the market if there exists a pair of dealers, j and k, who give the best bid and best ask respectively, and they fulfill the following condition,

$$B_j(t) \geq S_k(t)\,. \tag{42}$$

In the model the market price is given by the mean value of these two prices.

As a simple situation we consider a deterministic time evolution rule for these dealers. For all dealers the spread, $S_j(t) - B_j(t)$, is set to be a constant L. Each dealer has a position, either a seller or a buyer. When the jth dealer's position is a seller the selling price in mind, $S_j(t)$, decreases every time step until he can actually sell. Similar dynamics is applied to a buyer with the opposite direction of motion. In addition we assume that all dealers shift their prices in mind proportional to a market price change. When this proportional coefficient is positive, the dealer is categorized as a trend-follower. If this coefficient is negative, the dealer is called a contrarian. These rules are summarized by the following time evolution equations.

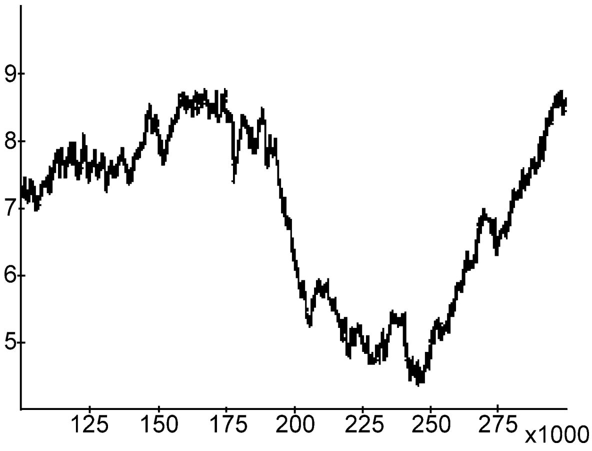
$$B_j(t+1) = B_j(t) + a_j S_j + b_j \Delta x(t)\,, \tag{43}$$

where S_j takes either +1 or −1 meaning the buyer position or seller position, respectively, $\Delta x(t)$ gives the latest market price change, $\{a_j\}$ are positive numbers given initially, $\{b_j\}$ are also parameters given initially.

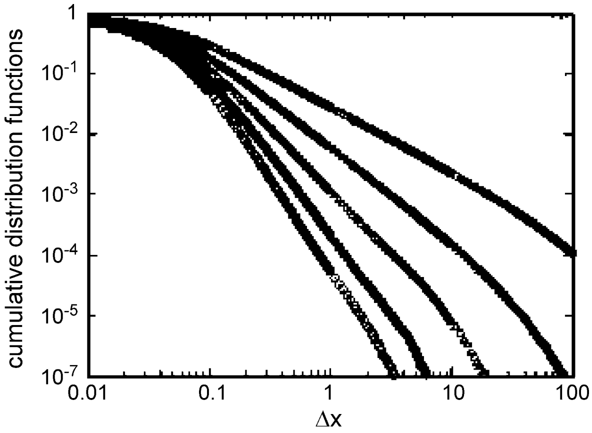
Figure 30 shows an example of market price evolution in the case of three dealers. It should be noted that although the system is deterministic, namely, the future price is determined uniquely by the set of initial values, resulting market price fluctuates almost randomly even in the minimum case of three dealers. The case of $N = 2$ gives only periodic time evolution as expected, while for $N \geq 3$ the system can produce market price fluctuations similar to the real market price fluctuations, for example, the fractal properties of price chart and the power law distribution of price difference are realized.

In the case that the value of $\{b_j\}$ are identical for all dealers, b, then the distribution of market price difference follows a power law where the exponent is controllable by this trend-follow parameter, b as shown in Fig. 31 [37]. The volatility clustering is also observed automatically for large dealer number case as shown in Fig. 32 (bottom) which looks quite similar to a real price difference time series Fig. 32 (top).

By adding a few features to this basic dealer model it is now possible to reproduce almost all statistical characteristics of market, such as tick-interval fluctuations, abnormal diffusions etc. [58]. In this sense the study of market behaviors are now available by computer simulations based on the dealer model. Experiments on the market is either impossible or very difficult for a real market, however, in an artificial market we can repeat occurrence of bubbles and crashes any times, so that we might be able to find a way to avoid catastrophic market behaviors by numerical simulation.



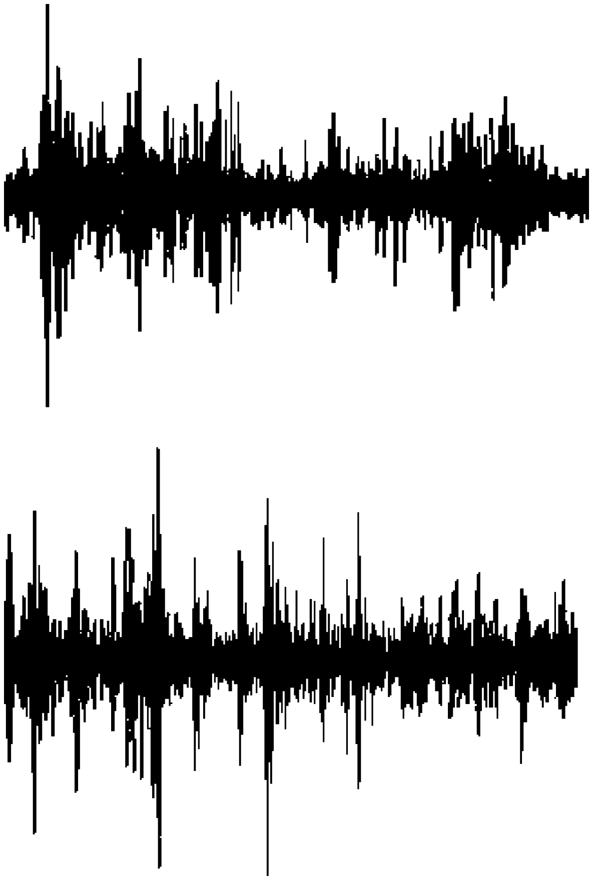
Fractals and Economics, Figure 30
Price evolution of a market with deterministic three dealers



Fractals and Economics, Figure 31
Cumulative distribution of a dealer model for different values of *b*. For weaker trend-follow the slope is steeper [38]

Income Distribution Models

Let us start with a famous historical problem, the St. Petersburg Paradox, as a model of income. This paradox was named after Daniel Bernoulli's paper written when he was staying in the Russian city, Saint Petersburg, in 1738 [6]. This paradox treats a simple lottery as described in the following, which is deeply related to the infinite expected



Fractals and Economics, Figure 32
Price difference time series for a real market (*top*) and a dealer model (*bottom*)

value problem in probability theory and also it has been attracting a lot of economists' interest in relation with the essential concept in economics, the utility [2].

Assume that you enjoy a game of chance, you pay a fixed fee, X dollars, to enter, and then you toss a fair coin repeatedly until a tail firstly appears. You win 2^n dollars where n is the number of heads. What is the fair price of the entrance fee, X?

Mathematically a fair price should be equal to the expectation value, therefore, it should be given as,

$$X = \sum_{n=0}^{\infty} 2^n \cdot \frac{1}{2^{n+1}} = \infty\,. \tag{44}$$

This mathematical answer implies that even X is one million dollars this lottery is generous enough and you should buy because expectation is infinity. But, would you dare to buy this lottery, in which you will win only one dollar with probability 0.5, and two dollars with probability 0.25, …?

Bernoulli's answer to this paradox is to introduce the human feeling of value, or utility, which is proportional to the logarithm of price, for example. Based on this expected utility hypothesis the fair value of X is given as follows,

$$X = \sum_{n=0}^{\infty} \frac{U(2^n)}{2^{n+1}} = \sum_{n=0}^{\infty} \frac{\log(2^n)}{2^{n+1}} = 1 + \log 2 \approx 1.69\,, \tag{45}$$

where the utility function, $U(x) = 1 + \log x$, is normalized to satisfy $U(1) = 1$. This result implies that the appropriate entry fee X should be about two dollars.

The idea of utility was highly developed in economics for description of human behavior, in the way that human preference is determined by maximal point of utility function, the physics concept of the variational principle applied to human action. Recently, in the field of behavioral finance which emerged from psychology the actual observation of human behaviors about money is the main task and the St. Petersburg paradox is attracting attention [36].

Although Bernoulli's solution may explain the human behavior, the fee $X = 2$ is obviously so small that the bookmaker of this lottery will bankrupt immediately if the entrance fee is actually fixed as two dollars and if a lot of people actually buy it. The paradox is still a paradox.

To clarify what is the problem we calculate the distribution of income of a gambler. As an income is 2^n with probability 2^{-n-1}, the cumulative distribution of income is readily obtained as,

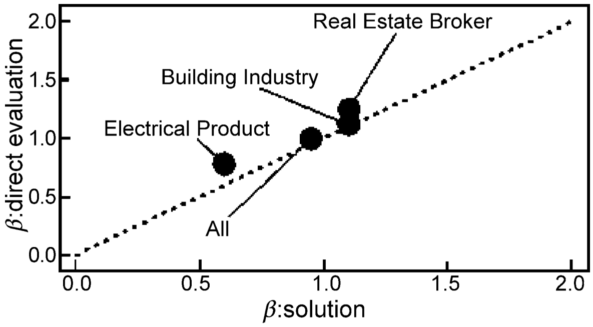
$$P(\geq x) \propto 1/x\,. \tag{46}$$

This is the power law which we observed for income distribution of companies in Sect. "Examples in Economics".

The key of this lottery is the mechanism that the prize money doubles at each time a head appears and the coin toss stops when a tail appears. By denoting the number of coin toss by t, we can introduce a stochastic process or a new lottery which is very much related to the St. Petersburg lottery.

$$x(t+1) = b(t)x(t) + 1\,, \tag{47}$$

where $b(t)$ is 2 with probability 0.5 and is 0 otherwise. As introduced in Sect. "Random Multiplicative Process", this problem is solved easily and it is confirmed that the steady state cumulative distribution of $x(t)$ also follows Eq. (46). The difference between the St. Petersburg lottery and the new lottery Eq. (47) is the way of payment of entrance fee. In the case of St. Petersburg lottery the entrance fee X is paid in advance, while in the case of new lottery you have to add one dollar each time you toss a coin. This new



Fractals and Economics, Figure 33
Theoretical predicted exponent value vs. observed value [29]

lottery is fair from both the gambler side and the bookmaker side because the expectation of income is given by $\langle x(t) \rangle = t$ and the amount of paid fee is also t.

Now we introduce a company's income model by generalizing this new fair lottery in the following way,

$$I(t+1) = b(t)I(t) + f(t)\,, \tag{48}$$

where $I(t)$ denotes the annual income of a company, $b(t)$ represents the growth rate which is given randomly from a growth rate distribution $g(b)$, and $f(t)$ is a random noise. Readily from the results of Sect. "Random Multiplicative Process", we have a condition to satisfy the empirical relation, Eq. (10),

$$\langle b(t) \rangle = \int bg(b) = 1\,. \tag{49}$$

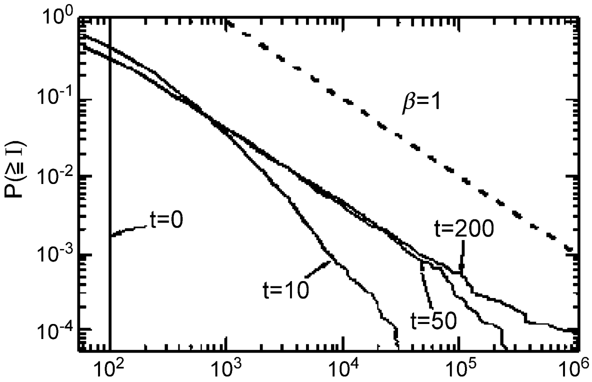
This relation is confirmed to hold approximately in actual company data [32].

In order to explain the job category dependence of the company's income distribution already shown in Fig. 12, we plot the comparison of exponents in Fig. 33. Empirically estimated exponents are plotted in the ordinate and the solutions of the following equation calculated in each job category are plotted in the abscissa,

$$\langle b(t)^{\beta} \rangle = 1\,. \tag{50}$$

The data points are roughly on a straight line demonstrating that the simple growth model of Eq. (48) seems to be meaningful.

An implication of this result is that if a job category is expanding, namely, $\langle b(t) \rangle > 1$, then the power law exponent determined by Eq. (50) is smaller than 1. On the other hand if a job category is shrinking, we have an exponent that is larger than 1.



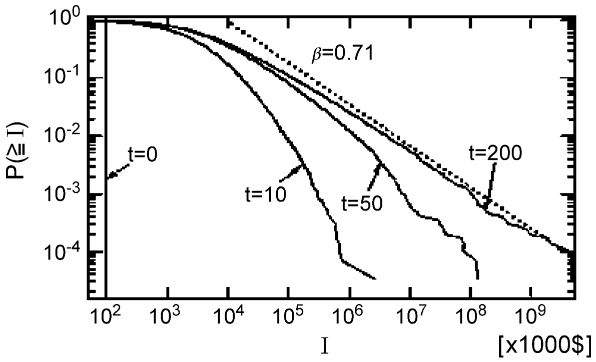
Fractals and Economics, Figure 34
Numerical simulation of income distribution evolution of Japanese companies [32]

This type of company's income model can be generalized to take into account the effect of company's size dependence on the distribution of growth rate. Also, the magnitude of the random force term can be estimated from the probability of occurrence of negative income. Then, assuming that the present growth rate distribution continues we can perform a numerical simulation of company's income distribution starting from a uniform distribution as shown in Fig. 34 for Japan and in Fig. 35 for USA. It is shown that in the case of Japan, the company size distribution converges to the power law with exponent −1 in 20 years, while in the case of USA the steady power law's slope is about −0.7 and it takes about 100 years to converge [31]. According to this result extremely large companies with size about 10 times bigger than the present biggest company will appear in USA in this century. Of course the growth rate distribution will change faster than this prediction, however, this model can tell the qualitative direction and the speed of change in very macroscopic economical conditions.

Other than this simple random multiplicative model approach there are various approaches to explain empirical facts about company's statistics assuming a hierarchical structure of organization, for example [23].

Future Directions

Fractal properties generally appear in almost any huge data in economics. As for financial market models, empirical fractal laws are reproduced and the frontier of study is now at the level of practical applications. However, there are more than a million markets in the world and little is known about their interaction. More research on market interaction will be promising. Company data



Fractals and Economics, Figure 35
Numerical simulation of income distribution evolution of USA companies [32]

so far analyzed show various fractal properties as introduced in Sect. "Examples in Economics", however, they are just a few cross-sections of global economics. Especially, companies' interaction data are inevitable to analyze the underlying network structures. Not only money flow data it will be very important to observe material flow data in manufacturing and consumption processes. From the viewpoint of environmental study, such material flow network will be of special importance in the near future. Detail sales data analysis is a new topic and progress is expected.

Bibliography

Primary Literature

1. Aoyama H, Nagahara Y, Okazaki MP, Souma W, Takayasu H, Takayasu M (2000) Pareto's law for income of individuals and debt of bankrupt companies. Fractals 8:293–300
2. Aumann RJ (1977) The St. Petersburg paradox: A discussion of some recent comments. J Econ Theory 14:443–445 http://en.wikipedia.org/wiki/Robert_Aumann
3. Bachelier L (1900) Theory of Speculation. In: Cootner PH (ed) The Random Character of Stock Market Prices. MIT Press, Cambridge (translated in English)
4. Bak P (1996) How Nature Works. In: The Science of Self-Organized Criticality. Springer, New York
5. Barabási AL, Réka A (1999) Emergence of scaling in random networks. Science 286:509–512; http://arxiv.org/abs/cond-mat/9910332
6. Bernoulli D (1738) Exposition of a New Theory on the Measurement of Risk; Translation in: (1954) Econometrica 22:22–36
7. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. J Political Econ 81:637–654
8. Brenner YS, Kaelble H, Thomas M (1991) Income Distribution in Historical Perspective. Cambridge University Press, Cambridge
9. Engle RF (1982) Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of UK Inflation. Econometrica 50:987–1008
10. Feller W (1971) An Introduction to Probability Theory and Its Applications, 2nd edn, vol 2. Wiley, New York
11. Friedlander SK (1977) Smoke, dust and haze: Fundamentals of aerosol behavior. Wiley-Interscience, New York
12. Fujiwara Y (2004) Zipf Law in Firms Bankruptcy. Phys A 337:219–230
13. Fujiwara Y, Guilmi CD, Aoyama H, Gallegati M, Souma W (2004) Do Pareto-Zipf and Gibrat laws hold true? An analysis with European firms. Phys A 335:197–216
14. Gilvarry JJ, Bergstrom BH (1961) Fracture of Brittle Solids. II Distribution Function for Fragment Size in Single Fracture (Experimental). J Appl Phys 32:400–410
15. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse Cubic Law for the Distribution of Stock Price Variations. Eur Phys J B 3:139–143
16. Handel PH (1975) 1/*f* Noise-An "Infrared" Phenomenon. Phys Rev Lett 34:1492–1495
17. Havlin S, Selinger RB, Schwartz M, Stanley HE, Bunde A (1988) Random Multiplicative Processes and Transport in Structures with Correlated Spatial Disorder. Phys Rev Lett 61: 1438–1441
18. Hidalgo CA, Klinger RB, Barabasi AL, Hausmann R (2007) The Product Space Conditions the Development of Nations. Science 317:482–487
19. Inaoka H, Toyosawa E, Takayasu H (1997) Aspect Ratio Dependence of Impact Fragmentation. Phys Rev Lett 78:3455–3458
20. Inaoka H, Ninomiya T, Taniguchi K, Shimizu T, Takayasu H (2004) Fractal Network derived from banking transaction – An analysis of network structures formed by financial institutions. Bank of Japan Working Paper. http://www.boj.or.jp/en/ronbun/04/data/wp04e04.pdf
21. Inaoka H, Takayasu H, Shimizu T, Ninomiya T, Taniguchi K (2004) Self-similarity of bank banking network. Phys A 339: 621–634
22. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2006) The Forbes 400 and the Pareto wealth distribution. Econ Lett 90:290–295
23. Lee Y, Amaral LAN, Canning D, Meyer M, Stanley HE (1998) Universal Features in the Growth Dynamics of Complex Organizations. Phys Rev Lett 81:3275–3278
24. Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–419
25. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, New York
26. Mandelbrot BB (2004) The (mis)behavior of markets. Basic Books, New York
27. Mantegna RN, Stanley HE (2000) An Introduction to Economics: Correlations and Complexity in Finance. Cambridge Univ Press, Cambridge
28. Markowitz HM (1952) Portfolio Selection. J Finance 7:77–91
29. Mizuno T, Katori M, Takayasu H, Takayasu M (2001) Statistical laws in the income of Japanese companies. In: Empirical Science of Financial Fluctuations. Springer, Tokyo, pp 321–330
30. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Analysis of high-resolution foreign exchange data of USD-JPY for 13 years. Phys A 324:296–302
31. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Investment strategy based on a company growth model. In: Takayasu H (ed) Application of Econophysics. Springer, Tokyo, pp 256–261

32. Mizuno T, Takayasu M, Takayasu H (2004) The mean-field approximation model of company's income growth. Phys A 332:403–411
33. Mizuno T, Toriyama M, Terano T, Takayasu M (2008) Pareto law of the expenditure of a person in convenience stores. Phys A 387:3931–3935
34. Ohnishi T, Takayasu H, Takayasu M (in preparation)
35. Okuyama K, Takayasu M, Takayasu H (1999) Zipf's law in income distribution of companies. Phys A 269:125–131. http://www.ingentaconnect.com/content/els/03784371;jsessionid=5e5wq937wfsqu.victoria
36. Rieger MO, Wang M (2006) Cumulative prospect theory and the St. Petersburg paradox. Econ Theory 28:665–679
37. Sato AH, Takayasu H (1998) Dynamic numerical models of stock market price: from microscopic determinism to macroscopic randomness. Phys A 250:231–252
38. Sato AH, Takayasu H, Sawada Y (2000) Power law fluctuation generator based on analog electrical circuit. Fractals 8:219–225
39. Sinha S, Pan RK (2008) How a "Hit" is Born: The Emergence of Popularity from the Dynamics of Collective Choice. http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.2955v1.pdf
40. Souma W (2001) Universal structure of the personal income distribution. Fractals 9:463–470; http://www.nslij-genetics.org/j/fractals.html
41. Takayasu H (1989) Steady-state distribution of generalized aggregation system with injection. Phys Rev Lett 63:2563–2566
42. Takayasu H (1990) Fractals in the physical sciences. Manchester University Press, Manchester
43. Takayasu H (ed) (2002) Empirical Science of Financial Fluctuations–The Advent of Econophysics. Springer, Tokyo
44. Takayasu H (ed) (2003) Application of Econophysics. Springer, Tokyo
45. Takayasu H, Inaoka H (1992) New type of self-organized criticality in a model of erosion. Phys Rev Lett 68:966–969
46. Takayasu H, Takayasu M, Provata A, Huber G (1991) Statistical properties of aggregation with injection. J Stat Phys 65:725–745
47. Takayasu H, Miura H, Hirabayashi T, Hamada K (1992) Statistical properties of deterministic threshold elements–The case of market price. Phys A 184:127–134
48. Takayasu H, Sato AH, Takayasu M (1997) Stable infinite variance fluctuations in randomly amplified Langevin systems. Phys Rev Lett 79:966–969
49. Takayasu M, Taguchi Y, Takayasu H (1994) Non-Gaussian distribution in random transport dynamics. Inter J Mod Phys B 8:3887–3961
50. Takayasu M (2003) Self-modulation processes in financial market. In: Takayasu H (ed) Application of Econophysics. Springer, Tokyo, pp 155–160
51. Takayasu M (2005) Dynamics Complexity in Internet Traffic. In: Kocarev K, Vatty G (eds) Complex Dynamics in Communication Networks. Springer, New York, pp 329–359
52. Takayasu M, Takayasu H (2003) Self-modulation processes and resulting generic 1/*f* fluctuations. Phys A 324:101–107
53. Takayasu M, Mizuno T, Takayasu H (2006) Potentials force observed in market dynamics. Phys A 370:91–97
54. Takayasu M, Mizuno T, Takayasu H (2007) Theoretical analysis of potential forces in markets. Phys A 383:115–119
55. Takayasu M, Mizuno T, Watanabe K, Takayasu H (preprint)
56. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. J Stat Phys 52:479–487; http://en.wikipedia.org/wiki/Boltzmann_entropy
57. Ueno H, Mizuno T, Takayasu M (2007) Analysis of Japanese bank's historical network. Phys A 383:164–168
58. Yamada K, Takayasu H, Takayasu M (2007) Characterization of foreign exchange market using the threshold-dealer-model. Phys A 382:340–346

Books and Reviews

Takayasu H (2006) Practical Fruits of Econophysics. Springer, Tokyo

Chatterjee A, Chakrabarti BK (2007) Econophysics of Markets and Business Networks (New Economic Windows). Springer, New York

GARCH Modeling

CHRISTIAN M. HAFNER
Université catholique de Louvain,
Louvain-la-Neuve, Belgium

Article Outline

Glossary

ACF Autocorrelation Function

ARMA Autoregressive Moving Average

BEKK A multivariate GARCH model named after an early unpublished paper by Baba, Engle, Kraft and Kroner.

CCC Constant Conditional Correlation

DCC Dynamic Conditional Correlation

CAPM Capital Asset Pricing Model

GARCH Generalized Autoregressive Conditional Heteroskedasticity

Heteroskedasticity A non-constant variance that depends on the observation or on time.

i.i.d. independent, identically distributed

Kurtosis A standardized fourth moment of a random variable that tells something about the shape of the distribution. A Gaussian distribution has a kurtosis of three. If the kurtosis is larger than three, then typically the distribution will have tails that are thicker than those of the Gaussian distribution.

Lag An operation that shifts the time index of a time series. For example, the first lag of y_t is y_{t-1}.

Long memory Property of covariance stationary processes without absolutely summable ACF, meaning that the ACF decays slowly.

Realized volatility Sum of intra-day squared returns as a measure for daily volatility.

Skewness A standardized third moment of a random variable that tells something about the asymmetry of the distribution. Symmetric distributions have skewness equal to zero.

Volatility Degree of fluctuation of a time series around its mean.

Definition of the Subject

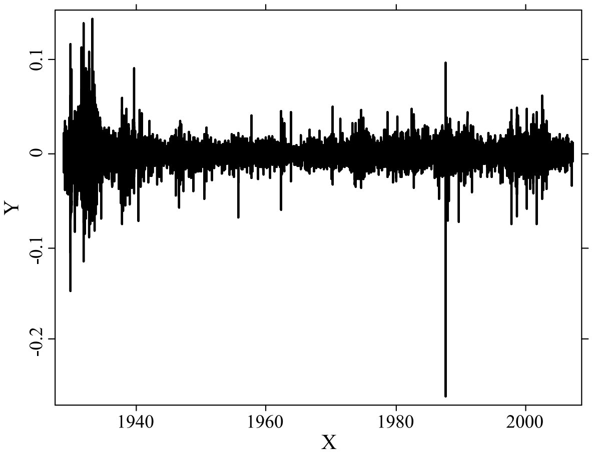
GARCH (Generalized Autoregressive Conditional Heteroskedasticity) is a time series model developed by [44] and [21] to describe the way volatility changes over time. In a GARCH model, the volatility at a given time t, σ_t^2 say, is a function of lagged values of the observed time series y_t. The GARCH model can be written as $y_t = \sigma_t \xi_t$, with ξ_t being an independent, identically distributed (i.i.d.) error term with mean zero and variance one, and where

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \qquad (1)$$

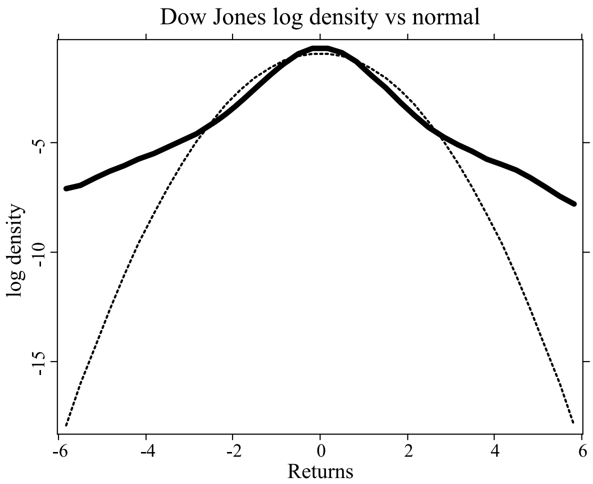
with constant parameters $\omega, \alpha_1, \ldots, \alpha_q$ and $\beta_1, \ldots, \beta_p$. Model (1) is also called GARCH(p,q), analogous to ARMA(p,q), as it includes p lagged volatilities and q lagged squared values of y_t. In this model, σ_t^2 is the variance of y_t conditional on the observations until time $t-1$. It is specified as a linear function of lagged squared y_t and lagged conditional variances. Many extensions or modifications of the basic model in (1) have been proposed, the most prominent being the exponential GARCH model of [99] and the threshold GARCH models of [123] and [59]. [71] and [42] provided classes of models that contain a large number of suggested models of the GARCH type.

Introduction

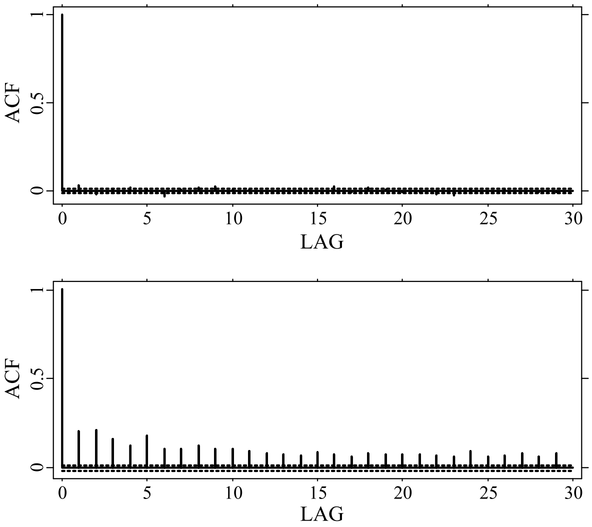
In the late seventies of the last century it became obvious that volatilities of financial assets are indeed not constant, nor deterministic or seasonal, but rather stochastic in nature. There is an unsystematic change between periods of high volatility and periods of low volatility. This 'volatility clustering' had already been remarked in the early works of [93] and [54]. It was one of several stylized facts of financial asset returns, another of which was the observation that the distribution of returns is not Gaussian. Of course, these features were not necessarily treated in an independent way, and in fact it was soon discovered that very likely one of the effects was causing another, such as volatility clustering causing leptokurtosis, or fat tailed distributions. For example, consider the simple model for as-



GARCH Modeling, Figure 1
Daily returns of the Dow Jones Index, 1928 to 2007, defined as first difference of the log index



GARCH Modeling, Figure 2
Dow Jones log density versus the Gaussian log density



GARCH Modeling, Figure 3
Dow Jones autocorrelation function of returns (*upper panel*) and squared returns (*lower panel*)

set returns y_t,

$$y_t = \sigma_t \xi_t$$

where $\xi_t \sim N(0,1)$ and σ_t is stochastic with $\mathrm{E}[\sigma_t^2] = \sigma^2$, say, and independent of present and future ξ_t. Then it is straightforward to show that the kurtosis of y_t is given by

$$\kappa = \frac{\mathrm{E}[y_t^4]}{\mathrm{E}[y_t^2]^2} = 3 + 3\frac{\mathrm{Var}(\sigma_t^2)}{\sigma^4}\,. \qquad (2)$$

Thus, returns in this model are Gaussian distributed if and only if $\mathrm{Var}(\sigma_t^2) = 0$, i. e., volatility is non-stochastic. Moreover, as the second term on the right hand side of (2) is always positive, the kurtosis will be larger than three under stochastic volatility, which often means that its tails are fatter than those of the Gaussian distribution. In other words, extreme events are more likely under stochastic volatility compared with constant volatility.

To illustrate the effects of volatility clustering and fat tails, consider the daily returns on the Dow Jones Industrial Index over the period October 1928 to April 2007. A graph of the (log) index X_t and returns, defined as $y_t = X_t - X_{t-1}$, is given in Fig. 1. Clearly visible is the volatility clustering in the beginning of the sample period and around the year 2000, while the years at the end of the sample showed less volatility. Also visible is the crash of October 1987 where the index dropped by 22 percent.

Figure 2 shows a nonparametric estimator of the logarithmic density of returns, compared with the analogue of a Gaussian distribution. Clearly, the Dow Jones returns distributions has fat tails, i. e., there are more extreme events than one would expect under normality. There are also more returns close to zero than under normality. Concerning volatility clustering, Fig. 3 shows the autocorrelation function of returns and squared returns. While there is very little structure in the ACF of returns, the ACF of squared returns are all positive and highly significant. This positive autocorrelation is explained by the fact that large returns tend to be followed by large returns and small returns tend to be followed by small returns.

Yet, realizing that volatility is stochastic does not tell us which model we should use for it. In practice, people are sometimes debating whether they should take historical volatilities over 20 or 100 days, say. They notice that calculating the standard deviation over a shorter period is more accurate when recent upturns or downturns want to be captured, while it is far less efficient than a longer time window when volatility has not changed much. Thus, there is some kind of bias-variance trade-off. The problem is that the optimal window length typically changes over time, and it is virtually impossible to adjust historical volatility windows automatically to market developments. A related problem is that historical volatilities imply a weighting scheme that is highly questionable: Why should k days be incorporated in the calculation with equal weight, but no weights are put to days up to $k+1$ days ago? A smoother weighting scheme seems more natural, and in particular, an exponential scheme seems attractive. Thus, for example, we may specify for σ_t^2

$$\sigma_t^2 = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i y_{t-1-i}^2 \tag{3}$$

with parameter $\lambda \in (0,1)$. Equation (3) can be rewritten as

$$\sigma_t^2 = (1-\lambda)y_{t-1}^2 + \lambda\sigma_{t-1}^2\,, \tag{4}$$

which looks more familiar. It is actually the model used by RiskMetrics of JP Morgan, when the smoothing parameter is fixed to 0.94. RiskMetrics is often used in practice as a means to calculate the Value-at-Risk (VaR) of a portfolio and to assess the market risk of a bank, required by the Basel Committee for Banking Supervision, see e. g., [78] and [95]. The VaR is essentially an extreme quantile of the distribution of portfolio returns. Under Gaussianity, for example, the VaR is a direct function of volatility. The RiskMetrics model is a special case of the integrated GARCH model of [47].

The generalized autoregressive conditional heteroskedasticity – GARCH – model is based on the seminal work of [44] and [21]. The idea is to do exponential smoothing in a more flexible way than RiskMetrics but keeping the model parsimonious. The particular specification reveals many similarities to autoregressive moving average (ARMA) time series models. In its most often used form, the standard GARCH model of order (1,1) reads

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{5}$$

where ω, α and β are parameters to be estimated from the data. Thus, the conditional variance is a linear function of lagged squared observations y_t and lagged conditional variances. Comparing (5) with the RiskMetrics model (4), it becomes clear that in the GARCH(1,1) model a constant is added, the parameter α takes the role of $1-\lambda$ and β that of λ. But since α and β can be chosen independently, the GARCH model is more flexible than RiskMetrics. In (5), substituting successively for σ_{t-i}^2, one obtains the analogue representation of (3),

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha\sum_{i=0}^{\infty}\beta^i y_{t-1-i}^2\,, \tag{6}$$

which clearly shows the exponential smoothing feature of the GARCH(1,1) model. The basic model can now be extended to allow for more lags. The GARCH(p,q) model is given by

$$\sigma_t^2 = \omega + \sum_{i=1}^{q}\alpha_i y_{t-i}^2 + \sum_{j=1}^{p}\beta_j\sigma_{t-j}^2 \tag{7}$$

extending the number of parameters to $p+q+1$. GARCH models of order higher than (1,1) allow for more complex autocorrelation structures of the squared process. However, in most empirical studies coefficients corresponding to higher lags turned out to be insignificant and thus simple GARCH(1,1) have clearly dominated models of higher order.

Although extremely successful due to its simplicity and yet accurate description of volatility changes, a thorough understanding of its stochastic properties such as stationarity or positivity constraints took many years. For example, [98] shows that, paradoxically at first sight, conditions for strict stationarity are less rigid than those for covariance stationarity if error terms are Gaussian, the reason being that covariance stationarity requires finite variances whereas strict stationarity does not. Moreover, the often given parameter restrictions $\omega > 0$, $\alpha_i, \beta_j \geq 0$ are only sufficient but not necessary for $\sigma_t > 0$ almost surely as demonstrated by [100]. These are just two examples for the subtleties of the theory of univariate GARCH processes.

Nevertheless, the immense success of the simple GARCH(1,1) model to explain many sorts of financial and macroeconomic time series was irreversible, partly also because it became available in standard statistical programming packages. The theory of estimation and inference developed rapidly, although perhaps still being underway, and estimation time of a GARCH(1,1) model for a thousand or so observations decreased from minutes in the eighties over seconds in the nineties to just fractions of a second nowadays. With these developments it became available to a broad public, and more and more practitioners started using the model, be it for option pricing, portfolio optimization, risk management, or other purposes. Monographs and reviews appeared such as [14,20,60] and [13]. Anniversary issues of renowned journals such as Journal of Applied Econometrics, 2002, were dedicated entirely to new ideas in GARCH modeling. The Nobel price for economics in 2003 was awarded to two time series econometricians, Clive Granger and Robert Engle. The latter has mainly driven the development of a new financial econometrics discipline, based on volatility modeling but spreading also to other areas such as modeling of extreme events and risk management.

The pricing of options and other derivatives is perhaps the most typical example for where models for the volatility of an asset matter. For example, the celebrated option pricing formula of [18] does not depend on the drift of the underlying stock but well on its volatility. In fact, among the ingredients of the Black and Scholes formula, volatility is the most crucial one, the other ones being either fixed such as time of maturity or strike price, or relatively easy to determine such as a riskfree interest rate. Volatility, however, has always been subject to debates about how exactly to find accurate measures for it. The Black and Scholes assumption of constant volatility is actually less crucial to their formula than one often thinks. Actually, if volatility is time-varying but in a deterministic way, then the Black and Scholes formula remains valid. One just has to replace the volatility parameter by the mean of the volatility function from today until the time of maturity of the option contract, see e. g., [90]. If, however, volatility is stochastic, i. e., it has an additional source of randomness, then markets are no longer complete and the Black and Scholes formula breaks down. In that case, assumptions about the volatility risk premium have to be made. In continuous time stochastic volatility models a classical paper is [74], while in a discrete time GARCH framework, [41] derives results for option pricing.

Properties of the GARCH(1,1) Model

For the sake of simplicity let us consider the univariate GARCH(1,1) model given in (5), where we additionally assume that the conditional distribution of y_t is Gaussian. The model can be written as

$$y_t = \sigma_t \xi_t , \quad \xi_t \sim \text{i.i.d. } N(0,1) \tag{8}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 . \tag{9}$$

In the following we discuss a few properties of model (8). First, the GARCH model specifies the *conditional* variance, where the condition is the information set generated by the process y_t. Formally, it is given by the sigma-algebra $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \ldots)$. With this notation we can write $\sigma_t^2 = \text{Var}(y_t|\mathcal{F}_{t-1})$, since σ_t^2 is $\mathcal{F}_{t-1}$-measurable. As the information set changes over time, the conditional variance also changes. On the other hand, this does not imply that the *unconditional* variance is also time-varying. In fact, for model (8) it is quite straightforward to show that the unconditional variance, if it exists, is constant and given by

$$\text{Var}(y_t) = \frac{\omega}{1-\alpha-\beta} .$$

A necessary and sufficient condition for the existence of the unconditional variance is

$$\alpha + \beta < 1 , \tag{10}$$

see [21]. He also shows that condition (10) is necessary and sufficient for the process $\{y_t\}$ to be covariance stationary. In that case, the autocorrelation function of $\{y_t\}$ is given by $\rho_\tau(y_t) = 0, \forall \tau \geq 1$. Moreover, both the conditional and unconditional mean of y_t are zero, so that the process $\{y_t\}$ has the properties of a white noise without being an i.i.d. process. The dependence occurs in higher moments of the process. For example, the autocorrelation function of the squared process, provided that fourth moments exist, is given by

$$\rho_1(y_t^2) = \alpha \frac{1-\alpha\beta-\beta^2}{1-2\alpha\beta-\beta^2} \tag{11}$$

$$\rho_\tau(y_t^2) = (\alpha+\beta)\rho_{\tau-1}(y_t^2), \quad \tau \geq 2 . \tag{12}$$

From (11) and (12) it is obvious that in the GARCH(1,1) model all autocorrelations of squared returns are positive with an exponential decay. This decay is slow if $\alpha + \beta$ is close to one, as often found for empirical data. One also characterizes this coefficient as the "persistence" parameter of the GARCH(1,1) model. The closer the persistence parameter is to one, the longer will be the periods of volatility clustering. On the other hand, the larger α relative to β, the higher will be the immediate impact of lagged squared returns on volatility.

The necessary and sufficient condition for finite fourth moments is given by $\beta^2 + 2\alpha\beta + 3\alpha^2 < 1$, see [21]. In that case, the kurtosis of y_t is given by

$$\kappa = 3 + \frac{6\alpha^2}{1-\beta^2-2\alpha\beta-3\alpha^2} ,$$

which is larger than three since the second term on the right hand side is positive. Hence, the GARCH(1,1) exhibits fat tails compared with a normal distribution.

A link between the GARCH model and an ARMA model is given by considering the squared process, $\{y_t\}$.

By simply adding y_t^2 and subtracting σ_t^2 on both sides of Eq. (9), one obtains

$$y_t^2 = \omega + (\alpha + \beta) y_{t-1}^2 - \beta u_{t-1} + u_t \tag{13}$$

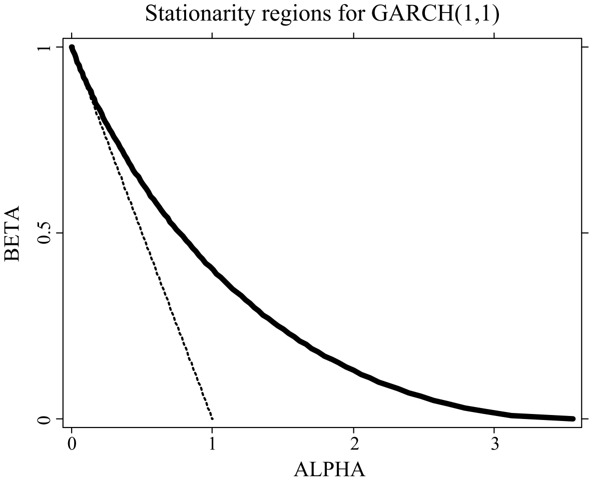
where $u_t = y_t^2 - \sigma_t^2$. Equation (13) is an ARMA(1,1) in y_t^2, since u_t is a white noise error term: We have $\mathrm{E}[u_t | \mathcal{F}_{t-1}] = 0$, which implies that all autocorrelations of u_t are zero.

It is possible that the process $\{y_t\}$ is strictly stationary without being covariance stationary, simply because a finite variance is not necessary for strict stationarity. If the process starts in the infinite past, a necessary and sufficient condition for strict stationarity of the GARCH(1,1) process as shown by [98] is given by

$$\mathrm{E}[\log(\alpha \xi_t^2 + \beta)] < 0 \,, \tag{14}$$

which is indeed weaker than condition (10). This follows directly by noting that (10) is equivalent to $\log(\alpha + \beta) = \log(\mathrm{E}[\alpha \xi_t^2 + \beta]) < 0$. Thus, by Jensen's inequality, $\mathrm{E}[\log(\alpha \xi_t^2 + \beta)] < \log(\mathrm{E}[\alpha \xi_t^2 + \beta]) < 0$. For example, for an ARCH(1) model (i. e., a GARCH(1,1) model with $\beta = 0$), α can be as large as 3.56 and still the process is strictly stationary. Figure 4 shows the different stationarity regions as a function of the two parameters. [25] generalized condition (14) to the GARCH(p, q) case.

Under the sufficient condition (10), the GARCH(1,1) process with Gaussian innovations is also geometrically ergodic and β-mixing with exponential decay as shown by [28].



GARCH Modeling, Figure 4

Stationarity regions for a GARCH(1,1) process with Gaussian innovations. To the *left* of the *dashed line* is the region of covariance stationarity, to the *left* of the *thick line* is the region of strict stationarity, and to the *right* of the *thick line* is the region of nonstationarity

If condition (14) holds, then the process $\{y_t\}$ has a stationary distribution whose tails are of the Pareto type. That is, for large x and some $a, k > 0$,

$$p(x) = Pr(y_t > x) = kx^{-a} \,. \tag{15}$$

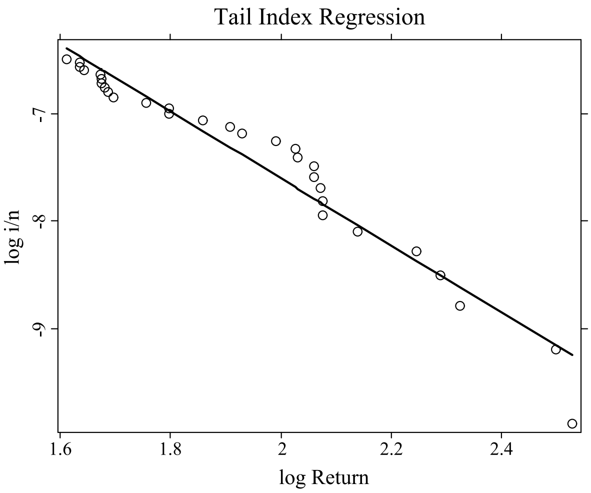
The coefficient a is known as the tail index. The smaller a, the fatter the tail of the distribution. For all c, $0 \leq c < \alpha$, $\mathrm{E}[|y_t|^c] < \infty$. [43] showed that a stationary ARCH model has Pareto-like tails. Knowledge of the tail index is important for risk management in order to assess the risk of extreme events. The theoretical tail index of a fitted ARCH or GARCH model can be compared with an estimate of the empirical tail index in order to diagnose the goodness-of-fit with respect to the tails. For example, taking logarithms of (15), one obtains $\log p(x) = \log(k) - a \log(x)$ for large x. Replacing x by the largest m order statistics of y_t, and introducing an error term, one obtains the regression

$$\log \frac{i}{n} = \log k - a \log X_{(i)} + \varepsilon_i \,, \quad i = 1, \ldots, m \tag{16}$$

where $X_{(i)}$ are the largest m order statistics of y_t and ε_i is an error term. One can estimate the coefficients of the linear regression (16) simply by ordinary least squares. More problematic is the choice of m, which involves a bias-variance trade-off. For the Dow Jones returns, Fig. 5 shows the tail index regression using $m = 30$. The OLS estimator of a is 3.12, indicating that fourth moments of returns may not exist. Another simple estimator is the Hill estimator proposed by [72], which is based on a likelihood principle. For the Dow Jones returns, the Hill estimator of a using $m = 30$ is 2.978, which is close to the OLS estimator, suggesting that even third moments may not exist. More elaborate estimators have been proposed and we refer to the detailed discussion in [43].

The presence of autoregressive conditional heteroskedasticity has an effect on the forecast intervals for predicted y_{t+k} given information at time t. If volatility at time t as measured by the GARCH model is high (low), these will be larger (smaller) than if GARCH effects are ignored. Furthermore, forecasting the volatility itself is easily possible with the standard GARCH model, since analytical expressions can be found for the conditional mean of future volatility as a function of today's information. The conditional mean is the optimal predictor in a mean square prediction error sense. For example, to forecast σ_{t+k}^2, one derives for a forecast horizon of $k \geq 2$,

$$\mathrm{E}[\sigma_{t+k}^2 | \mathcal{F}_t] = \omega(1 + (\alpha + \beta) + \cdots + (\alpha + \beta)^{k-2}) \\ + (\alpha + \beta)^{k-1} \sigma_{t+1}^2 \,.$$



GARCH Modeling, Figure 5
Tail index regression for the Dow Jones returns

If the process is covariance stationary, i. e., $\alpha + \beta < 1$, then volatility forecasts converge to the unconditional variance:

$$\lim_{k\to\infty} \mathrm{E}[\sigma^2_{t+k}|\mathcal{F}_t] = \frac{\omega}{1-\alpha-\beta} = \mathrm{Var}(y_t)\,.$$

In the early literature on GARCH models, these were criticized for not providing good forecasts in terms of conventional forecast criteria. For example, when regressing the ex post squared daily returns on the forecasted conditional variance, the obtained R^2 is typically small, of the order of about ten percent. [4] found that the daily squared return is not really the targeted value, but that daily volatility should rather be measured by the sum of *intra-day* squared returns, e. g., on intervals of five minute returns, which they called *realized volatility*. In terms of realized volatility, the forecasting performance of GARCH models improved substantially to levels of about fifty percent R^2. Later, a new branch of volatility modeling opened by noticing that if intra-day data are available, then it is indeed more efficient to measure daily volatility directly by realized volatility and then do forecasting of daily volatility using models fitted to realized volatility, see e. g., [5].

Estimation and Inference

The principal estimation method for GARCH models is maximum likelihood (ML). In most cases one assumes a conditional Gaussian distribution. If the true distribution is Gaussian, then ML estimators are consistent and efficient under quite general conditions. On the other hand, if the true distribution is not Gaussian, then one loses efficiency but again under quite general conditions, consistency is retained if at least the first two conditional moments are correctly specified, see [23]. In the case of misspecification of the conditional distribution one also speaks of *quasi* maximum likelihood (QML), distinguishing it from ML where the true distribution is used, which however in general is unknown.

The log likelihood function, up to an additive constant and conditional on some starting value for the volatility process, reads $L(\theta) = \sum_{t=1}^n l_t(\theta)$, where

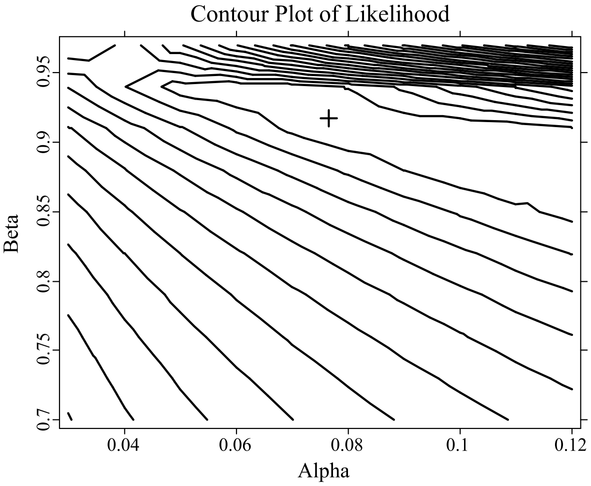
$$l_t(\theta) = -\frac{1}{2}\log\sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^n \frac{y_t^2}{\sigma_t^2(\theta)}$$

and where $\theta = (\omega, \alpha, \beta)'$ is the parameter vector. The maximum likelihood estimator is then defined as the maximizer of $L(\theta)$ over some compact set Θ,

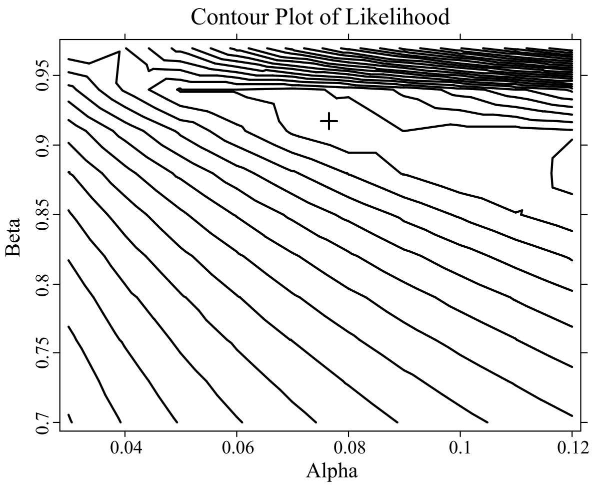
$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta)\,. \qquad (17)$$

Unfortunately, there is no closed form solution to (17) but many numerical optimization procedures exist. For example, a popular algorithm is that of [15].

Figure 6 shows the likelihood function of a GARCH(1,1) process generated using the parameter estimates of the Dow Jones index returns (see Sect. "Asymmetry, Long Memory, GARCH-in-Mean"), Gaussian innovations, and the same sample size of $n = 19727$. The parameter ω has been determined by the variance targeting technique of [50], i. e., $\omega = \sigma^2(1 - \alpha - \beta)$, where σ^2 is



GARCH Modeling, Figure 6
Contour plot of the likelihood function of a generated GARCH(1,1) process using Gaussian innovations and a sample size of $n = 19727$. The abscissa is the parameter α, the ordinate is β. True values, marked by a *cross* in the figure, are $\alpha = 0.0766$ and $\beta = 0.9173$



GARCH Modeling, Figure 7
Contour plot of the likelihood function of the GARCH(1,1) model fitted to observed Dow Jones index returns, 1928 to 2007, with sample size $n = 19727$. The *abscissa* is the parameter α, the *ordinate* is β. The maximum, marked by a *cross* in the figure, is obtained for $\alpha = 0.0766$ and $\beta = 0.9173$

the sample variance of observed returns. In regions where $\alpha + \beta \geq 1$, ω is set to zero. Note the steep decline of the likelihood for values of α and β that lie beyond the covariance stationarity region ($\alpha + \beta \geq 1$). Figure 7 shows the same function for the observed Dow Jones index returns. No major difference can be detected between both graphs, indicating an appropriate specification of the Gaussian likelihood function.

If the first two moments of y_t are correctly specified and under further regularity conditions given by [118] and [23], the QML estimator is consistent with asymptotic distribution given by

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}) \tag{18}$$

where

$$\mathcal{I} = \mathrm{E}\left[\frac{\partial l_t}{\partial \theta}\frac{\partial l_t}{\partial \theta'}\right], \quad \mathcal{J} = -\mathrm{E}\left[\frac{\partial^2 l_t}{\partial \theta \partial \theta'}\right],$$

and where the derivatives are evaluated at the true parameter values. In case the conditional distribution is indeed Gaussian, one has the identity $\mathcal{I} = \mathcal{J}$ and the asymptotic covariance matrix reduces to the inverse of the information matrix, $\mathcal{I}^{-1}$. Note that consistency is retained if the conditional distribution is not Gaussian, but efficiency is lost in that case.

It is straightforward to obtain analytical formula for the score vector, the outer product of the score and the Hessian matrix, with which inference on parameter estimates can be done using the result in (18). More primitive conditions than those of [23] have been derived, e. g., by [57,85,91] and [66].

Maximum likelihood estimation using other than Gaussian distributions has been considered, e. g., by [101]. He shows that if the distribution is misspecified, then consistency is no longer guaranteed. In particular, if a symmetric distribution is assumed but the true distribution is asymmetric, then maximum likelihood estimators are inconsistent. In practice, a common distribution used for maximum likelihood estimation is the Student t distribution. Given the results of [101], one should be careful in interpreting parameter estimates if there is evidence for skewness in standardized residuals.

Another estimation strategy based on maximum likelihood is a nonparametric estimation of the error density, which has been advocated by [48]. They suggest to use a first stage estimator of the model parameters, which is consistent but not efficient such as the Gaussian MLE, to construct residuals and then to use nonparametric methods to estimate the error density. Given the estimated error density, one can maximize the likelihood corresponding to this nonparametric density function. These estimators will under regulatory conditions be consistent and more efficient than the Gaussian ML estimator, provided that the true density is different from Gaussian.

A potential practical problem of maximum likelihood estimators is its dependence on numerical optimization routines. Recently, a closed form estimator based on the autocorrelation structure of squared returns has been suggested by [82]. Their estimator is inefficient compared to ML but has the advantage of being uniquely determined by the data. Further Monte Carlo evidence is necessary to see whether it is a serious practical competitor for ML-type estimators. Least squares type estimators of ARCH(q) have been considered by [118] and [103]. Again, these are inefficient compared with maximum likelihood estimators but simpler to compute. [104] suggest a least absolute deviation estimator for GARCH models that is robust with respect to outliers but does not allow for a closed form. Finally, Bayesian estimation of GARCH-type models has been investigated, e. g., by [12,115] and [32].

Testing for ARCH

In a regression such as

$$y_t = \mu_t + \varepsilon_t \tag{19}$$

where μ_t is measurable w.r.t. $\mathcal{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, \ldots)$ and ε_t is a white noise sequence, inference on μ_t typically depends on the properties of the error term ε_t. For example, if ε_t is i.i.d. Gaussian and μ_t is linear such as an AR(p) model, then estimation by least squares of the autoregressive coefficients in μ_t is efficient. If, however, ε_t is not i.i.d. and for example conditionally heteroskedastic, then estimation by ordinary least square (OLS) is no longer efficient and some kind of generalized least squares may be employed. Moreover, inference on the parameters in μ_t will be erroneous if homoskedasticity of ε_t is assumed but, in reality, ε_t is conditionally heteroskedastic. In particular, standard errors in that case are typically underestimated. To avoid this, it is essential to test for ARCH type effects in ε_t. The following testing procedure, based on the Lagrange multiplier principle, has been proposed in the original ARCH paper by [44]. The null hypothesis is that ε_t is i.i.d. white noise, the alternative is the presence of ARCH. One first estimated the model (19) by least squares, obtains residuals $\hat{\varepsilon}_t$, and then runs the regression

$$\hat{\varepsilon}_t^2 = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \alpha_2 \hat{\varepsilon}_{t-2}^2 + \cdots + \alpha_q \hat{\varepsilon}_{t-q}^2 + \eta_t \quad (20)$$

where η_t is an error term. Under the null hypothesis H_0, $\alpha_1 = \ldots = \alpha_q = 0$. The test statistic is $\lambda = nR^2$, where n is the sample size and R^2 the coefficient of determination of the regression (20). Under H_0, the test statistic follows asymptotically a χ^2 distribution with q degrees of freedom. Hence, it is an elementary exercise to test for ARCH effects in the error term of regression models. Historically, it is remarkable that prior to the introduction of the ARCH model, the above LM test was used by Prof. Clive Granger as an LM test for a bilinear error term, for which it has some power. Then, Prof. Robert Engle discovered that it has more power for another model, which he then introduced as the ARCH model.

An alternative to the LM test of [44] would be a Wald-type test of the hypothesis $H_0\colon \alpha = 0$ in the GARCH(1,1) model (5) using, e. g., the t-ratio as test statistic. However, this test is non-standard since under the null hypothesis the parameter α is on the boundary of the parameter space and the parameter β is not identified. [6] treats this test in a general framework.

Asymmetry, Long Memory, GARCH-in-Mean

In the standard GARCH model in (7), positive and negative values of lagged returns y_{t-i} have the same impact on volatility, since they appear in squares in the equation for σ_t^2. Empirically, it has been frequently noted since [17] that for stock markets, negative returns increase volatility more than positive returns do. Essentially, this so-called leverage effect means that negative news have a stronger impact on volatility than positive ones. To account for this empirical observation, several extensions of the standard GARCH model have been proposed in the literature. The most commonly used are the exponential GARCH model of [99] and the threshold GARCH model of [59] and [123]. The threshold model in its first order variant is given by the process

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \alpha^* y_{t-1}^2 I(y_{t-1} < 0) + \beta \sigma_{t-1}^2$$

where α^* is an additional parameter and $I(\cdot)$ is the indicator function. If $\alpha^* = 0$, then the threshold model reduces to the standard GARCH model. If $\alpha^* > 0$, then negative returns have a stronger impact on volatility than positive ones, which corresponds to the empirical observation for stock markets.

Secondly, the exponential GARCH (EGARCH) model of [99] specifies log-volatility as

$$\log \sigma_t^2 = \omega + \theta \xi_{t-1} + \alpha(|\xi_{t-1}| - \mathrm{E}|\xi_{t-1}|) + \beta \log \sigma_{t-1}^2$$

where $\xi_t = y_t/\sigma_t$ is i.i.d. with a generalized error distribution (GED) which nests the Gaussian and allows for slightly fatter tails. Due to the specification of log-volatility, no parameter restrictions are necessary to keep volatility positive. Moreover, the conditions for weak and strong stationarity coincide. Note that if $\theta \neq 0$, then $\mathrm{Cov}(y_t^2, y_{t-j}) \neq 0$ such that a leverage effect can be captured. A drawback of the EGARCH model is that asymptotic theory for maximum likelihood estimation and inference under primitive conditions are not available yet, but [110] are making much progress in this respect.

Another model allowing for asymmetry is the asymmetric power ARCH (APARCH) model [36]. In its (1,1) order form it specifies volatility as

$$\sigma_t^\delta = \omega + \alpha(|y_{t-1}| - \phi y_{t-1})^\delta + \beta \sigma_{t-1}^\delta$$

where δ is a positive parameter. If $\delta = 2$ and $\phi = 0$, the standard GARCH model is retained. For $\phi \neq 0$, there is an asymmetric impact of positive and negative lagged returns on volatility. The additional flexibility due to the parameter δ allows to better reproduce the so-called 'Taylor property,' originally noted by [111], which says that the autocorrelations of $|y_t|^d$ are positive even at long lags, and when viewed as a function of d take a maximum for $d \approx 1$ for many financial returns y_t. [70] provide a formal discussion of this issue.

The standard GARCH(p, q) model (7) implies that the decay of the autocorrelation function (ACF) of squared returns is geometrically fast. However, one often finds evidence for a slow hyperbolical decay in financial time series, see for example Fig. 3. The decay pattern of the ACF is related to the structure of coefficients c_j in the ARCH(∞) representation of GARCH models,

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j y_{t-j}^2 \,. \tag{21}$$

For example, in the GARCH(1,1) model, these are given by $c_j = \alpha\beta^{j-1}$. Covariance stationary GARCH models have the property that the autocovariance function of squared returns, $\gamma(\tau) = \mathrm{Cov}(y_t^2, y_{t-\tau}^2)$, is absolutely summable, i. e., $\sum_\tau |\gamma(\tau)| < \infty$. Such a property is commonly called *short memory* as opposed to *long memory* processes for which the ACF is not absolutely summable. Long memory GARCH models have been proposed by [8] and [19], see also the review of long memory processes in econometrics by [7]. An example of a long memory GARCH process would be given by (21) with $c_j = Cj^{-\theta}$ for some constant C and parameter $\theta > 0$. A particular example for such a process is the fractionally integrated GARCH (FIGARCH) model of [8], which can be written as

$$(1-L)^d \sigma_t^2 = \omega + \alpha y_{t-1}^2$$

where L is the lag operator and d a positive parameter. When $d = 1$ one obtains the integrated GARCH (IGARCH) model of [47]. For $d \neq 1$ one can use a binomial extension to obtain after inverting

$$(1-L)^{-d} = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} L^j = \sum_{j=0}^{\infty} c_j L^j \tag{22}$$

where $\Gamma(\cdot)$ is the Gamma function. The coefficient c_j in (22) can be shown to be of the long memory type. A similar long memory EGARCH model has been introduced by [19]. The drawback of these particular specifications is that they share the property with the IGARCH model to have infinite variance. [105] has proposed a long memory GARCH type model that allows for finite variance.

Finally, in the finance literature a link is often made between the expected return and the risk of an asset, since investors are willing to hold risky assets only if their expected returns compensate for the risk. A model that incorporates this link is the GARCH-in-mean or GARCH-M model of [52], given by

$$y_t = \delta g(\sigma_t^2) + \varepsilon_t$$

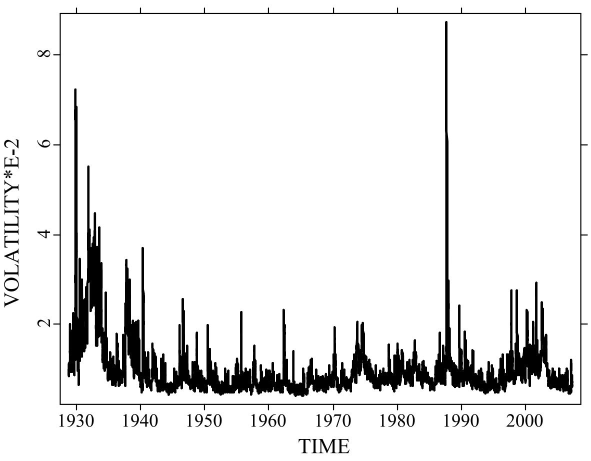
where ε_t is an ARCH or GARCH error process, δ a parameter, and g a known function such as square root or logarithm. If $\delta > 0$ and g is monotone increasing, then the term $\delta g(\sigma_t^2)$ can be interpreted as a risk premium that increases expected returns $\mathrm{E}[y_t]$ if volatility σ_t^2 is high. It can be shown that such a model, when applied to the market index, is consistent with the capital asset pricing model (CAPM) of [108] and [87], see [24].

As an empirical illustration we estimate alternative models for the Dow Jones index discussed in the introduction. To recall, we have daily returns from October 1928 to April 2007. First order autocorrelation of returns is 0.03, which due to the large number of observations is significant at the level 1%. However, we refrain here from fitting an autoregressive or moving average model to the returns as the results concerning volatility estimation do not change substantially. We only consider a constant conditional mean in the model $y_t = \mu + \varepsilon_t$, where ε_t is one of the discussed GARCH-type models and μ takes into account a non-zero trend in returns. Six alternative GARCH models are considered, all of them being of order (1,1): standard GARCH, TGARCH, EGARCH, GARCH-M, TGARCH-M and EGARCH-M. For the "in-mean" versions, we have chosen the square root specification for the function $g(\cdot)$, which seems to work better than the logarithm or the identity function. Moreover, for all "in-mean" models the constant μ turned out to be insignificant and hence was suppressed from the model. Table 1 summarizes the estimation results.

Note first that all estimated risk premia are positive $\mu > 0$ and $\delta > 0$, as theory would predict. Second, for all models allowing for asymmetry, the leverage effect of negative returns is confirmed, i. e., $\alpha^* > 0$ for the TGARCH models and $\theta < 0$ for the EGARCH models. Third, in all cases persistence of shocks to volatility is very high, measured by $\alpha + \beta$ in the GARCH model, $\alpha + \alpha^*/2 + \beta$ in the TGARCH model (assuming a symmetric innovation distribution), and by β in the EGARCH model. Thus, all models are short memory with exponential decay of the ACF of squared returns, but the models try to adapt to the empirically observed slow decay of the ACF by pushing the persistence parameter close to one. This near-IGARCH behavior is typical for daily returns. Finally, the goodness-of-fit seems to be best for the TGARCH-in-mean model, taking the log-likelihood as criterion. The estimation results strongly confirm the presence of the leverage effect, high persistence, and positive risk premium in the data. Figure 8 shows the estimated conditional standard deviation of the TGARCH-M model. For the other models, the graph would look quite similar and is therefore not shown here. Notice the very volatile periods at the beginning of the sample in the 1930s, around the year 2000 corresponding to the "new economy" boom and following crash, as well as the spike in 1987 due to the crash of October 17, 1987.

GARCH Modeling, Table 1
Estimation results for the following models: GARCH, EGARCH, TGARCH, GARCH-in-mean, TGARCH-in-mean and EGARCH-in-mean, applied to daily Dow Jones returns from 1928 to 2007. All parameters are significant at the one percent level

	G	TG	EG	GM	TGM	EGM
μ	4.25E-04	2.64E-04	2.36E-04			
δ				0.0591	0.0354	0.0286
ω	8.74E-07	1.03E-06	-0.2184	8.75E-07	1.06E-06	-0.2227
α	0.0766	0.0308	0.1391	0.0766	0.0306	0.1378
$\alpha*$		0.0769			0.0761	
θ			-0.0599			-0.0595
β	0.9173	0.9208	0.9879	0.9172	0.9205	0.9874
L	65456.47	65600.53	65589.32	65462.45	65603.79	65590.69



GARCH Modeling, Figure 8
Estimated conditional standard deviation of daily Dow Jones index returns, 1928 to 2007, using the TGARCH-in-mean model

Non- and Semi-parametric Models

Nonparametric methods refrain from associating particular parametric forms to functions or distributions. Instead, only the class of functions is determined, for example the class of squared integrable functions or the degree of smoothness. The price for the flexibility is typically slower convergence rates than parametric models. A combination of the two approaches is often called semiparametric. One such approach has already been mentioned in Sect. "Estimation and Inference" in the context of estimation by maximum likelihood using nonparametric estimates of the error density, as proposed by [48] for GARCH models. [88] shows that this procedure leads to adaptive estimation of the identifiable parameters (α and β in a GARCH(1,1) model) in the sense of [16]. That is, it is possible to achieve the Cramer–Rao lower bound and do as good as if one knew the true error distribution. The scale parameter ω, however, is not adaptively estimable. See also [40] and [37] for related results for univariate GARCH models, and [65] for an extension to semiparametric estimation of multivariate GARCH models.

A different approach is to directly model the volatility process in a nonparametric way. Early models were proposed by [61] and [51]. The qualitative threshold ARCH model of [61] specifies models of the type

$$y_t = \sum_{j=1}^{J} \sigma_j I(y_{t-1} \in A_j)\xi_t \,, \tag{23}$$

where (A_j) is a partition of the real line and σ_j, $j = 1, \dots, J$, are positive parameters. Thus, volatility is modeled as a piecewise constant function of lagged returns. Note that the threshold ARCH model of [59] and [123] is not a special case of (23) as there the volatility function is piecewise quadratic in lagged returns. Extensions to the ARCH(q) and GARCH(p, q) are straightforward. [51] replaced the piecewise constant functions by piecewise linear functions. In both cases, one may consider their models as nonparametric if the partition becomes finer as the sample size increases.

Consider the model

$$y_t = \sigma(y_{t-1})\xi_t$$

where $\sigma(\cdot)$ is an unknown smooth function, and $\xi_t \sim$ i.i.d. $N(0,1)$. For $\sigma^2(x) = \alpha x^2$ one obtains the parametric ARCH(1) model of [44]. An example of a nonparametric estimator of $\sigma(\cdot)$ is the Nadaraya–Watson estimator given by

$$\hat{\sigma}^2(x) = \frac{\sum_{t=2}^{n} K\{(y_{t-1} - x)/h\} y_t^2}{\sum_{t=2}^{n} K\{(y_{t-1} - x)/h\}}$$

where K is a kernel function satisfying $\int K(x)\mathrm{d}x = 1$ and $\int xK(x)\mathrm{d}x = 0$, and where $h > 0$ is a bandwidth that determines the degree of smoothing. As the Nadaraya–Watson estimator can be interpreted as fitting a constant locally, a generalization consists of fitting a local polynomial instead. This has been derived by [68] for the volatility case.

A general problem of nonparametric methods is the so-called *curse of dimensionality* when smoothing has to operate in high dimensions. Considering a nonparametric ARCH(q) model,

$$y_t = \sigma(y_{t-1}, \ldots, y_{t-q})\xi_t$$

this problem is apparent and in practice very large data sets are required to estimate the function g with appropriate precision. One may be inclined to impose more structure on the g function such as additive or multiplicative separability. Nonparametric multiplicative ARCH models have been proposed by [63] and [120]. Semi-parametric additive ARCH models of the type $\sigma^2(y_{t-1}, \ldots, y_{t-q}) = \sum_{j=1}^{p} \beta^{j-1} g(y_{t-j})$ with some unknown function g and parameter $\beta \in (0,1)$ have been considered by [29].

Extension of nonparametric ARCH(q) models to nonparametric GARCH(1,1) models have also been proposed. However, in its general form $y_t = \sigma_t \xi_t$ with $\sigma_t = g(y_{t-1}, \sigma_{t-1})$, the model is difficult to estimate due to lack of structure. One might consider iterative estimation algorithms, based on some initial estimate of volatility as in [26].

Imposing a similar semi-parametric structure as for the semi-parametric ARCH(q) model of [29], one can write $\sigma_t^2 = g(y_{t-1}) + \beta\sigma_{t-1}^2$, where again $g(\cdot)$ is an unknown smooth function. Note that this model nests many of the proposed parametric models. It has been considered by [119] and [89].

In practice, nonparametric methods may be used whenever it is not a priori clear what functional form fits best the data, either by using them directly, or as a tool to specify a parametric model in a second stage.

Multivariate GARCH Models

In economics and finance, one typically deals with multiple time series that are fluctuating in a non-systematic manner and are considered as realizations of stochastic processes. The interest for applied econometricians is therefore to model their risk, that is, their volatility, but also their inter-dependencies. For example, if one has reasons to assume that the underlying stochastic processes are Gaussian, then the inter-dependencies may be completely described by the correlation structure. In fact, when we say 'multivariate volatility models' we usually mean the modeling of volatilities but also that of correlations. This is also the reason why the extension of univariate volatility to multivariate volatility models is much more complex than that of univariate models for the conditional mean, such as ARMA models, to the multivariate case.

It will be immediately clear that the multivariate case is the one that is by far more relevant in practice when financial markets are under study. The reason is, first, the large number of different assets, or even different types of contracts, assets, exchange rates, interest rates, options, futures, etc. Second, there is usually a strong link between these variables, at least within one group. For example, asset returns in one stock market tend to be quite strongly correlated. One would make big approximation errors when treating the variables as independent by, e. g., using univariate volatility models for the conditional variances and set conditional covariances to zero. Note that, setting conditional covariance to zero is much stronger an assumption than setting the unconditional covariance to zero. Ways must be found to treat the dependence of the series in a flexible yet parsimonious way.

A first step would again be to do exponential smoothing à la RiskMetrics, which can be used not only to obtain the individual variances according to (4), but also to obtain the correlations. To see this, we define just as in (4) an exponential smoother for the covariances as

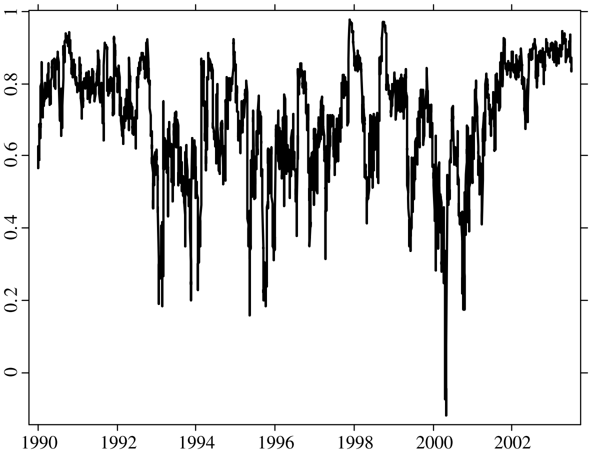
$$\sigma_{12,t} = (1-\lambda)\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \lambda\sigma_{12,t-1}$$

and then obtain as usual the conditional correlation as

$$\rho_t = \frac{\sigma_{12,t}}{\sigma_{1,t}\sigma_{2,t}} \,,$$

which is guaranteed to be between minus one and one if the same parameter λ is used, typically $\lambda = 0.94$. Figure 9 depicts the RiskMetrics conditional correlation series for the DOW and NASDAQ return series.

Obviously, conditional correlations are not constant, although it is difficult from the graph to verify such a statement statistically. However, one thing to observe is that during the New Economy boom in 1999 and 2000, estimated correlations have been substantially lower, sometimes even negative, than at other times. The reason is probably a decoupling due to the higher vulnerability of the NASDAQ with respect to the bubble in high tech and internet stocks. A more thorough analysis of this data set which also compares this model with other, more flexible models is provided by [45]. We see that the RiskMetrics tool, even though very simple, can give some guidelines. One of the objectives of the econometrician is to enhance the model in terms of flexibility (e. g., why should λ be fixed to 0.94?), and to establish a statistical framework



GARCH Modeling, Figure 9
Conditional correlations of the Dow Jones IA and NASDAQ index returns, daily, using the RiskMetrics model

in which hypotheses such as constant conditional correlations can be tested.

From an econometrical viewpoint, modeling the volatility of multiple time series is, for several reasons, challenging both theoretically and practically. For the sake of illustration, consider a bivariate GARCH model of the Vec type that was introduced by [24]. Denote by H_t the conditional variance matrix of the asset return vector y_t. Then a bivariate ARCH(1) model reads

$$H_t = \begin{pmatrix} h_{1,t} & h_{12,t} \\ h_{12,t} & h_{2,t} \end{pmatrix}$$

and where

$$\begin{aligned} h_{1t} &= \omega_1 + \alpha_{11}\varepsilon^2_{1,t-1} + \alpha_{12}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{13}\varepsilon^2_{2,t-1} \\ h_{12,t} &= \omega_2 + \alpha_{21}\varepsilon^2_{1,t-1} + \alpha_{22}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{23}\varepsilon^2_{2,t-1} \\ h_{2t} &= \omega_3 + \alpha_{31}\varepsilon^2_{1,t-1} + \alpha_{32}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{33}\varepsilon^2_{2,t-1} \,. \end{aligned}$$

Each conditional variance, h_{1t} and h_{2t}, and conditional covariance, $h_{12,t}$, depends on all lagged squared returns (two in the bivariate case) and all lagged cross-products (one in the bivariate case). The main reason for the rapidly increasing complexity of the model when the dimension is increased lies in the fact that not only all conditional variances with their cross-dependencies have to be modeled, but also all conditional correlations. It is in fact the latter that poses the main problem, as there are a total of $N(N-1)/2$ such correlations when the dimension is N, but only N variances. Thus, modeling variances and correlations simultaneously, a total of $N(N+1)/2$ entries of the conditional covariance matrix need to be modeled. For example, if $N = 10$ (a moderate dimension for many economic or financial problems) this number is 55, if $N = 100$ (such as modeling all stocks of a common stock index), then 5050 series, conditional variances and covariances, are at stake.

It is clear that this is too much to allow for a flexible cross-dependence of the individual series. Without imposing any structure except linearity, the multivariate generalization of the standard GARCH model, the so-called Vec model introduced by [24], is feasible only for low dimensions, two or three say, as otherwise the number of parameters becomes too high relative to the number of observations typically encountered in economic practice. Another problem is that the Vec model does not guarantee a positive definite covariance matrix. Necessary conditions for the latter desirable property are as yet unknown in the general Vec specification.

These are some reasons to look for other models, and in fact, over recent years a broad variety of different approaches to the problem have been suggested in the literature. Roughly speaking, one can divide them into two groups. The first one tries to simplify the problem by imposing more structure on the Vec model. Examples are the BEKK model of [49] and the factor GARCH model by [53]. More recently, the second group tries to separate the problem of modeling the conditional variances and conditional correlations. An early and simple version of this group is to say that conditional variances are just univariate GARCH and conditional correlations are constant over time, as suggested by [22]. In its simplicity, this constant conditional correlation (CCC) model basically does not add any complexity beyond univariate GARCH to the multivariate estimation problem, which renders the model extremely useful in empirical practice. It also introduced the idea of two-step estimation, where in the first step conditional variances are modeled, and in the second step the conditional correlations using the standardized residuals of the first step. However, starting with [45] there have been plenty of arguments in favor of time-varying conditional correlations in financial markets. In particular, a common finding is that correlations are higher when the market moves up than when it moves down. A test for this correlation asymmetry has been suggested by [73]. Using a dynamic conditional correlation model (DCC), [45] shows that time varying correlations are not uncommon even in normal market situations. In the following we sketch these two branches of the multivariate GARCH literature. It should however be mentioned that there are models that do not fall into these two categories such as a multivariate version of the exponential GARCH model proposed by [79].

Factor GARCH Models

In the following factor GARCH models are discussed as an example of multivariate GARCH models. The main idea of factor models is to reduce the dimension of the system to a tractable two or three factors, which can then be modeled in a standard way. It should be noted that also 'full-factor' models with number of factors equal to the number of variables have been proposed in the literature. For example, [116] propose the model

$$y_t = W f_t$$

where W is a $N \times N$ parameter matrix and f_t is a N-vector with conditional mean zero and diagonal conditional variance matrix, Σ_t say. The individual conditional variances of f_t can be modeled by univariate GARCH(1,1), for example. One can restrict W to be lower triangular, as it is well known that the Choleski decomposition of a positive definite matrix always exists and is unique. Thus, the conditional variance matrix of y_t is given by $H_t = W \Sigma_t W = L_t L_t'$, where $L_t = W \Sigma_t^{1/2}$ is lower triangular. In this model, the parameters in W and those in Σ_t need to be estimated jointly, which may be cumbersome in high dimensions. The empirical performance of such full factor models still remains to be investigated.

It is more common to specify only a few factors and allow for idiosyncratic noise. We will look at such models in the following. Suppose that there are K (observed or unobserved) factors, collected in a K-vector f_t, with $K < N$. Then a simple factor model can be written as

$$y_t = W f_t + v_t \tag{24}$$

where v_t is a white noise vector with $\text{Var}(v_t) = \Omega$ that represents the idiosyncratic noise. Typically, one assumes that Ω is diagonal so that components of the idiosyncratic noise are uncorrelated. In that case, correlation between components of y_t is induced only through the common factors f_t. If y_t represents the error of a time series system, one may constrain f_t to have conditional mean zero. The matrix W is of dimension $N \times K$, of full column rank, and contains the so-called *factor loadings*, the weights of a factor associated with the individual components of y_t.

In finance, model (24) is well known from the arbitrage pricing theory (APT) of [106], where y_t are excess returns of financial assets, f_t are systematic risk factors and v_t is unsystematic risk. It can also be viewed as a generalization of the capital asset pricing model (CAPM) developed by [108] and [87]. For simplicity we assume here that factors are observed. If they are unobserved, identification issues arise that are discussed, e. g., by [107].

For the factors, a low-dimensional GARCH model can be assumed: $\text{Var}(f_t \mid \mathcal{F}_{t-1}) = \Sigma_t$, where Σ_t is a $(K \times K)$ covariance matrix. The conditional covariance matrix of y_t is given by

$$H_t = \text{Var}(y_t \mid \mathcal{F}_{t-1}) = W \Sigma_t W' + \Omega\,. \tag{25}$$

In the case of just one factor, the matrix W reduces to a vector w and the factor volatility, σ_t^2 say, can be modeled by univariate GARCH and the conditional variance of y_t simplifies to

$$H_t = ww'\sigma_t^2 + \Omega\,.$$

If the factors are conditionally uncorrelated, i. e., Σ_t is diagonal with $\Sigma_t = \text{diag}(\sigma_{1t}^2, \ldots, \sigma_{Kt}^2)$, then one can write

$$H_t = \sum_{k=1}^{K} w_k w_k' \sigma_{kt}^2 + \Omega$$

where w_k is the kth column of W. [83] propose methods to test for the number of factors K and derive results for maximum likelihood estimation. For the more general BEKK model class, [31] derived asymptotic theory but assuming moments of order eight of the process, which may exclude many of the typically fat-tailed financial time series.

A popular factor GARCH model is the orthogonal GARCH (OGARCH) model of [3]. In the OGARCH model, factors f_t are the K largest principal components obtained from the (unconditional) sample covariance matrix, and the loading matrix W is the matrix of associated eigenvectors. The loadings represent the sensitivity of an individual series on a specific factor. By construction, the unconditional correlation between the factors is zero, due to the orthogonality of the principal components. However, the *conditional* correlation may be different from zero. Denote the (empirical) covariance matrix of y_t by Σ. The decomposition $\Sigma = \Gamma \Lambda \Gamma'$ gives $\Gamma = (\gamma_1, \ldots, \gamma_N)$ with the eigenvectors γ_i and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$ with corresponding eigenvalues λ_i. We order the columns of Γ according to the magnitude of the corresponding eigenvalues such that $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Let us assume here that all eigenvalues are distinct, otherwise Γ may not be identified. For the case of non-distinct eigenvalues, one may use the more general singular value decomposition and go for the GO-GARCH (generalized orthogonal GARCH) model of [114].

The vector of principal components, given by

$$\begin{pmatrix} f_t \\ \varepsilon_t \end{pmatrix} = \Gamma' y_t$$

is partitioned into the first K components f_t, whose volatility will assumed to be stochastic, and the last $N-K$ components ε_t, whose volatility will assumed to be constant. One could speak of K *dynamic* and $N-K$ *static* factors.

Now decompose the matrices as follows:

$$\Gamma_{(N\times N)} = (\Gamma_{1_{(N\times K)}}, \Gamma_{2_{(N\times(N-K))}}) \,,$$

and $\Lambda_1 = \operatorname{diag}(\lambda_1,\dots,\lambda_K)$, $\Lambda_2 = \operatorname{diag}(\lambda_{K+1},\dots,\lambda_N)$.

The model can then be written as

$$y_t = \Gamma_1 f_t + \Gamma_2 \varepsilon_t \tag{26}$$

where $\operatorname{Var}(f_t \mid \mathcal{F}_{t-1}) = \Sigma_t$ and $\operatorname{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \Lambda_2$. For example, Σ_t may be diagonal or some K-variate GARCH model. Note that this representation is equivalent to that of (24) with $W = \Gamma_1$ and $v_t = \Gamma_2\varepsilon_t$, except that $\Omega = \Gamma_2\Lambda_2\Gamma_2'$ will not be diagonal in general. The conditional variance of y_t is given by

$$\begin{aligned} H_t &= \operatorname{Var}(y_t \mid \mathcal{F}_{t-1}) = \Gamma_1\Sigma_t\Gamma_1' + \Gamma_2\Lambda_2\Gamma_2' \\ &= \Gamma \begin{pmatrix} \Sigma_t & 0 \\ 0 & \Lambda_2 \end{pmatrix} \Gamma' \,. \end{aligned}$$

If Σ_t follows a K-variate BEKK process, then it can be shown that H_t will follow an N-variate BEKK process with restrictions on the parameter matrices. However, the classical OGARCH assumes that factors are conditionally orthogonal, hence Σ_t is diagonal, additional to the fact that they are unconditionally orthogonal by construction. This assumption is crucial and may not always be justified in practice. It should be emphasized that in the OGARCH model, the factor loadings contained in the matrix Γ and the factor variances contained in Λ are considered as fixed for a given sample covariance matrix. This contrasts the general factor model (24) where factor loadings W are estimated jointly with the parameters describing the factor dynamics.

Instead of using unconditionally orthogonal factors, [55] proposed to use conditionally orthogonal factors by searching numerically for linear combinations of the data such that the conditional correlation between these combinations is minimized under norm constraints. The existence of such linear combinations is tested using bootstrap methods.

Constant and Dynamic Conditional Correlation Models

[22] suggests a multivariate GARCH model with constant conditional correlations. Let H_t be the conditional covariance matrix of a series y_t, and V_t be a diagonal matrix with the conditional standard deviations of y_t on its diagonal. Then the model is simply

$$H_t = V_t R V_t \tag{27}$$

where R is the constant correlation matrix. H_t is positive definite as long as the conditional variances are positive and R is positive definite. For instance, one could specify univariate GARCH models for the individual conditional variances. One the other hand, it is possible to allow for spill-over of volatilities from one series to other series. Note that the CCC model is not nested in the Vec specification. Theory of maximum likelihood estimation for CCC-type models has been established by [77] for consistency and [86] for asymptotic normality.

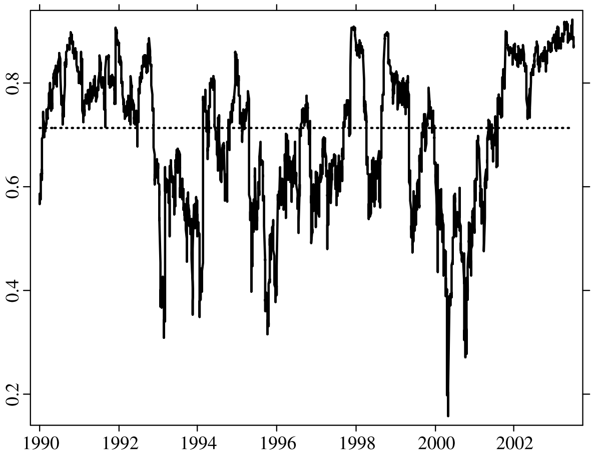
The assumption of constant correlations simplifies strongly the estimation problem. However, it might sometimes be too restrictive. For example, it is often observed that correlations between financial time series increase in turbulent periods, and are very high in crash situations. A Lagrange Multiplier test against the CCC model has been suggested by [112]. An extension of the CCC model to allow for time-varying correlations is the dynamic conditional correlations (DCC) model introduced by [45]. The DCC model renders the conditional correlation matrix R dependent on time, R_t say. The conditional correlation between the ith and jth component of y_t is modeled as

$$R_{ij,t} = \frac{Q_{ij,t}}{\sqrt{Q_{ii,t}Q_{jj,t}}}$$

where $Q_{ij,t}$ is the ijth element of the matrix Q_t given by

$$Q_t = S(1-\alpha-\beta) + \alpha v_{t-1}v_{t-1}' + \beta Q_{t-1} \tag{28}$$

where α and β are parameters and $v_t = V_t^{-1}y_t$ are the standardized but correlated residuals. That is, the conditional variances of the components of v_t are one, but the conditional correlations are given by R_t. The matrix S is the sample correlation matrix of v_t, so a consistent estimate of the unconditional correlation matrix. If α and β are zero, we get the above CCC model. If they are different from zero one gets a kind of ARMA structure for all correlations. Note however that all correlations would follow the same kind of dynamics, since the ARMA parameters are the same for all correlations. The specification of the first term of Q_t ensures that the unconditional mean of Q_t is equal to the sample covariance matrix of v_t, similar to the *variance targeting* technique of [50]. Also it facilitates the estimation, since that can be done in two steps: First, the conditional variances in V_t are estimated using univariate GARCH models, for example, then v_t, the



GARCH Modeling, Figure 10
Conditional correlations of the Dow Jones IA and NASDAQ index returns, daily, using the DCC model. *Dashed line*: constant conditional correlation

standardized (but correlated) residuals and their covariance matrix S are computed, before in the second step only two remaining parameters, α and β, need to be estimated. A model similar to DCC has been proposed by [113].

Figure 10 depicts the estimated conditional correlations for the DOW Jones and NASDAQ time series, using the DCC and CCC models. Comparing the former with the RiskMetrics estimates of Fig. 9, no substantial difference can be detected visually. However, the parameter estimates of α and β are 0.0322 and 0.9541 with standard errors 0.0064 and 0.0109, respectively, so that the null hypothesis $H_0\colon \alpha = 0.06$ is clearly rejected. Whether or not the difference in estimated conditional correlations matters in empirical applications has been addresses, e. g., by [30], who consider the problem of portfolio selection. [96] compare the performance of CCC, DCC, OGARCH and a model of [84] in forecasting and portfolio selection in high dimensions. They find that the difference is not substantial, but that the CCC model is too restrictive.

To summarize, the whole challenge of multivariate volatility modeling is to balance model complexity and simplicity in such a way that the model is flexible enough to capture all stylized facts in the second moments (and perhaps beyond that) of the series while keeping it simple for estimation and inference.

In the following we sketch some applications of multivariate GARCH models in finance. As an early example, [24] estimate a capital asset pricing model (CAPM) with time-varying betas. The beta is defined as the ratio of the asset return's covariance with the market return, divided by the variance of the market return. Denote by r_{it} the excess return of asset i, and by r_{mt} the excess return of the market. Then the beta-form of the CAPM can be written as

$$r_{it} = \beta_{it} r_{mt} + \varepsilon_i = \frac{\mathrm{Cov}(r_{it}, r_{mt})}{\mathrm{Var}(r_{mt})} + \varepsilon_i$$

where ε_i is idiosyncratic noise whose risk cannot be diversified away and is therefore called unsystematic risk. As we observe time varying second moments, it is clear that betas will also be time varying, not only due to the variance of the market but also due to the covariances of the assets with the market. However, if both returns are covariance stationary, then by definition the unconditional second moments will be constant, and only after conditioning on suitable information sets such as historical returns will second moments become time varying.

Secondly, correlations between exchange rates have been substantially time-varying, as for example in Europe the European exchange rate mechanism enforced increasing correlations. The correlation of the DEM/USD and FRF/USD rates, for instance, increased steadily in the late 1990s until it was virtually one just before the launch of the Euro. See, e. g., [45], who models these data, among others, with alternative correlation models. Thirdly, portfolio selection is another type of application. If, for example, one is interested in the minimum variance portfolio of n assets with covariance matrix Σ, then the well known formula for the optimal weight vector α is given by

$$\alpha = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota}$$

where ι is an n-vector of ones, see, e. g., [30]. Obviously, if Σ is allowed to be time-varying, then the optimal portfolio weights will in general also depend on time. This has many important practical implications, e. g., for portfolio managers. One of the problems is to determine an optimal reallocation frequency. If variances and covariances change daily and the objective is to minimize the portfolio variance over the next ten days, then one could follow at least two strategies: either calculate the optimal portfolio weights daily and reallocate accordingly. Or, calculate the return distribution over ten days, obtain thus a covariance matrix for ten-day returns, find the optimal weights using this covariance matrix and leave the corresponding portfolio unchanged throughout the ten days. If the objective is to minimize the variance over the ten days, then the first method will usually outperform the second. The intuitive reason is that the second method aggregates data, thus losing valuable information. However, in practice one may still prefer the second method for various reasons, one of which could be the higher transaction costs of the first method.

Stochastic Volatility

GARCH models discussed so far explain the conditional variance at time t as a function of the information set at time $t-1$. In other words, it is measurable with respect to this information set. This is not the case for models of the stochastic volatility (SV) type, which introduce an extra error term in the volatility equation. For example, in the univariate case such a model could take the form

$$\begin{aligned} y_t &= \sigma_t \xi_t \\ \log \sigma_t^2 &= \omega + \beta \log \sigma_{t-1}^2 + \eta_t \end{aligned} \tag{29}$$

where ξ_t and η_t are i.i.d. mean zero random variables with variance equal to one and σ_η^2, respectively. Here, log volatility follows an AR(1) process. Since volatility is unobserved, model (29) is a particular case of a latent variable model. Note that, if the information set at time $t-1$ consists of all lagged values of y_t up to y_{t-1}, then volatility at time t is not measurable with respect to this information set. [27] compare moment properties such as kurtosis and persistence of SV and GARCH models. [58] propose a model that encompasses both GARCH and stochastic volatility and thus allows for testing against each of them.

Estimation is more complicated than for GARCH models because the likelihood is an integral of dimension equal to the sample size, given by

$$L(Y;\theta) = \int p(Y \mid H, \theta) p(H \mid \theta) \mathrm{d}H \tag{30}$$

where $Y = (y_1, \ldots, y_n)$, $H = (\sigma_1^2, \ldots, \sigma_n^2)$, and $\theta = (\omega, \beta, \sigma_\eta^2)$. Maximization of (30) has no closed form and numerical optimization is difficult due to the high dimensional integral. Therefore, other estimation methods have been considered in the literature, for example generalized method of moments (GMM), simulated maximum likelihood with Markov Chain Monte Carlo (MCMC) or Bayesian methods, see e. g., [75,76] and [80]. An application to currency options by [92] compares three alternative estimation algorithms and finds that the estimation error of the volatility series is large for all methods.

In the multivariate case, without imposing structure, estimating a highly dimensional stochastic volatility model seems difficult. One way of imposing structure in multivariate SV models is to assume a factor model as, e. g., in [34,69,81] and [56], or constant correlations. To consider a bivariate extension of stochastic volatility models, one suggestion of [69] is to say that the stochastic variances $\sigma_{1,t}$ and $\sigma_{2,t}$ of the two assets follow univariate stochastic variance processes as in (29), and the stochastic covariance is given by

$$\sigma_{12,t} = \rho \sigma_{1,t} \sigma_{2,t} \,,$$

where ρ is a constant parameter between -1 and 1. This model, very much in the spirit of the constant conditional correlation GARCH model of [22], is quite parsimonious and can be efficiently estimated using simulated maximum likelihood as demonstrated in [33]. It is straightforward to generalize this specification to higher dimensions. However, estimation may then become trickier. Also the restriction of constant correlation parameters may not be innocuous. More empirical tests are required about goodness of fits, comparing the non-nested GARCH and SV type models of about the same model complexity.

SV models lend themselves naturally to continuous time stochastic volatility models and realized volatility. Indeed, as shown by [9], realized volatility can be used to estimate the volatility of SV models. The monograph of [109] collects influential papers of the stochastic volatility literature.

Aggregation

The frequency at which financial time series are sampled is often not unique. For example, one researcher may be interested in the behavior of returns to the Dow Jones index at a daily frequency, but another one at a weekly or monthly frequency. Considering log-returns, weekly returns can be directly obtained from daily returns by simply summing up intra-week returns. If a model is fitted to daily returns, an important question is what this implies for the weekly returns. In particular, one may ask if the model remains in the same class, which would then be called closed under temporal aggregation. For the univariate GARCH model, [38] have shown that only a weak version of it is closed under temporal aggregation. Instead of modeling the conditional variance, weak GARCH models the best linear predictor of squared returns in terms of a constant, lagged returns and lagged squared returns. In the weak GARCH(1,1) case, they show how to obtain the parameters of the aggregated process as a function of the parameters of the high frequency process. In particular, denoting the parameters of the aggregated process by $\alpha^{(m)}$ and $\beta^{(m)}$, where m is the aggregation level, then the persistence parameter of the aggregated level is given by $\alpha^{(m)} + \beta^{(m)} = (\alpha + \beta)^m$. Thus, the persistence of the aggregated process declines geometrically fast with the aggregation level. Asymptotically, the process will reduce to white noise. One would therefore expect to see much less conditional heteroskedasticity in monthly returns than in weekly or daily returns. The link between parameters at different frequencies also provides a means for model diagnostics. The results of [38] have been extended to the multivariate case by [64].

Instead of aggregating, one could go the other way and look at "disggregating" the process temporally, i. e., sampling the underlying process at finer intervals. [97] showed that GARCH models can be viewed as approximations of continuous time stochastic volatility models, see also [39]. However, [117] has shown that the GARCH model and its diffusion limit are not equivalent in a statistical experiment sense.

Rather than aggregating temporally, one may alternatively be interested in aggregating contemporaneously in a multivariate context. For example, stock indices are constructed as linear combinations of individual stocks. [102] show that again the aggregated process is only weak GARCH. Rather than aggregating multivariate GARCH models, one can alternatively consider aggregation of univariate heterogenous GARCH processes with random coefficients. In linear ARMA models, this aggregation scheme is known to produce long memory type behavior of the aggregate, see [62]. [35] conjectured that this holds in a similar way for GARCH models. However, [122] shows that although the ACF of the squared aggregate decays hyperbolically, it may be absolutely summable and hence there is no long memory. For the general model class of [94] which includes GARCH, weak GARCH and stochastic volatility as special cases, [121] shows that contemporaneous aggregation leads to long memory properties of the aggregate.

Future Directions

The theory of univariate GARCH models is now well developed and understood. For example, theory of maximum likelihood estimation is available under weak conditions that allow for integrated and even mildly explosive processes. However, theory of multivariate GARCH is still in its infancy and far from closed, due to arising technical difficulties. For general specification such as the BEKK model, no results on asymptotic normality of estimates are available yet that would allow for integrated processes. Most available results on general specifications are high level and only for some special cases, primitive conditions are established. This is certainly one of the main directions for future research.

On the modeling side, there is no clear general answer how to deal with the problem of high dimensions, and in particular how to balance model flexibility with econometric feasibility. More practical experience is necessary to see what type of model performs best for what kind of data. On the application side, a still open issue is how to evaluate the volatility risk for option pricing, and how to efficiently use multivariate GARCH models in portfolio selection or risk management. Other frontiers for GARCH models are discussed by [46].

An interesting new field is the combination of GARCH models with nonparametric distributions to obtain more accurate estimates of the Value-at-Risk, mentioned in Sect. "Introduction". In the univariate case this is quite obvious, but in the multivariate case one has to deal with the "curse of dimensionality", common in the nonparametrics literature. Furthermore, issues such as tail dependence need to be modeled accurately in that case. A joint framework that captures volatilities, correlations, other distributional shape features and tail dependence would be an interesting target for applied research.

Finally, *realized volatilities* (RV) have been mentioned at the end of Sect. "Properties of the GARCH(1,1) Model" as a means to use intra-day data to generate accurate ex post measures of daily volatilities. Using these RV measures, one can build time series models that predict daily volatilities one or more steps ahead, see e. g., [5] for a detailed analysis. It seems that RV provides better forecasts than GARCH, which is not surprising as it uses more information, namely the intra-day returns. The RV literature has evolved as an important second branch of volatility modeling next to the discrete time GARCH or SV models. One direction of research is the treatment of microstructure noise, present in most high frequency data, as e. g., in [1,2] and [67]. Another one is the modeling of jumps using the so-called bi-power variation and the generalization to the multivariate case using realized covariances and bipower co-variation, see e. g., [10] and [11]. Other directions are possible and it seems likely that RV will become the dominant econometric tool to model volatilities provided that high frequency data are available.

Bibliography

Primary Literature

1. Ait-Sahalia Y, Mykland P, Zhang L (2005) A tale of two time scales: Determining integrated volatility with noisy high-frequency data. J Am Stat Assoc 100:1394–1411
2. Ait-Sahalia Y, Mykland P, Zhang L (2005) How often to sample a continuous time process in the presence of market microstructure noise. Rev Financial Stud 18:351–416
3. Alexander C (2001) Orthogonal GARCH. In: Mastering Risk, Financial Times, vol 2. Prentice Hall, London, pp 21–38
4. Andersen TG, Bollerslev T (1998) Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. Int Econ Rev 39:885–905
5. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71:579–625
6. Andrews DWK (1999) Estimation when a parameter is on a boundary. Econometrica 67:1341–1383

7. Baillie RT (1996) Long memory processes and fractional integration in econometrics. J Econom 73:5–59
8. Baillie RT, Bollerslev T, Mikkelsen HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. J Econom 74:3–30
9. Barndorff-Nielsen O, Shephard N (2002) Econometric analysis of realized volatility and its use in estimating stochastic volatility models. J Roy Stat Soc 64:253–280
10. Barndorff-Nielsen O, Shephard N (2004) Econometric analysis of realized covariation: high frequency covariance, regression and correlation in financial economics. Econometrica 72:885–925
11. Barndorff-Nielsen O, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps (with discussion). J Financial Econom 2:1–48
12. Bauwens L, Lubrano M (1998) Bayesian inference on garch models using the gibbs sampler. Econom J 1:C23–C46
13. Bauwens L, Laurent S, Rombouts J (2006) Multivariate garch models: a survey. J Appl Econom 21:79–109
14. Bera A, Higgins M (1993) A survey of ARCH models: properties, estimation and testing. J Econ Surv 7:305–366
15. Berndt EK, Hall BH, Hall RE, Hausman JA (1974) Estimation and inference in nonlinear structural models. Ann Econ Soc Meas 3:653–665
16. Bickel PJ (1982) On adaptive estimation. Ann Stat 10:647–671
17. Black F (1976) Studies in stock price volatility changes. In: Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association, pp 177–181
18. Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Political Econ 81:637–654
19. Bollerslev T, Mikkelsen HO (1996) Modeling and pricing long-memory in stock market volatility. J Econom 73:151–184
20. Bollerslev T, Engle R, Nelson D (1994) ARCH models. In: Engle R, McFadden D (eds) Handbook of Econometrics. North Holland, Amsterdam, pp 2959–3038
21. Bollerslev TP (1986) Generalized autoregressive conditional heteroscedasticity. J Econom 31:307–327
22. Bollerslev TP (1990) Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. Rev Econ Stat 72:498–505
23. Bollerslev TP, Wooldridge JM (1992) Quasi maximum likelihood estimation of dynamic models with time-varying covariances. Econom Rev 11:143–172
24. Bollerslev TP, Engle RF, Wooldridge JM (1988) A capital asset pricing model with time-varying covariances. J Political Econ 96:116–131
25. Bougerol P, Picard N (1992) Stationarity of garch processes and some nonnegative time series. J Econom 52:115–127
26. Bühlmann P, McNeil AJ (2002) An algorithm for nonparametric GARCH modelling. Comput Stat Data Anal 40:665–683
27. Carnero MA, Pena D, Ruiz E (2004) Persistence and kurtosis in GARCH and stochastic volatility models. J Financial Econom 2:319–342
28. Carrasco M, Chen X (2002) Mixing and moment properties of various garch and stochastic volatility models. Econom Theory 18:17–39
29. Carroll R, Härdle W, Mammen E (2002) Estimation in an additive model when the components are linked parametrically. Econom Theory 18:886–912
30. Chan LKC, Karceski J, Lakonishok J (1999) On portfolio optimization: Forecasting covariances and choosing the risk model. Rev Financial Stud 12:937–674
31. Comte F, Lieberman O (2003) Asymptotic theory for multivariate garch processes. J Multivar Anal 84:61–84
32. Concepcion Ausian M, Galeano P (2007) Bayesian estimation of the gaussian mixture GARCH model. Comput Stat Data Anal 51:2636–2652
33. Danielsson J (1998) Multivariate stochastic volatility models: Estimation and a comparison with vgarch models. J Empir Finance 5:155–174
34. Diebold FX, Nerlove M (1989) The dynamics of exchange rate volatility: A multivariate latent factor arch model. J Appl Econom 4:1–21
35. Ding Z, Granger C (1996) Modelling volatility persistence of speculative returns: a new approach. J Econom 73:185–215
36. Ding Z, Granger CWJ, Engle RF (1993) A long memory property of stock market returns and a new model. J Empir Finance 1:83–106
37. Drost FC, Klaassen CAJ (1997) Efficient estimation in semiparametric garch models. J Econom 81:193–221
38. Drost FC, Nijman T (1993) Temporal aggregation of GARCH processes. Econometrica 61:909–927
39. Drost FC, Werker BJM (1996) Closing the garch gap: Continuous garch modeling. J Econom 74:31–57
40. Drost FC, Klaassen CAJ, Werker BJM (1997) Adaptive estimation in time series models. Ann Stat 25:786–817
41. Duan JC (1995) The GARCH option pricing model. Math Finance 5:13–32
42. Duan JC (1997) Augmented garch(p,q) process and its diffusion limit. J Econom 79:97–127
43. Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events. Springer, Berlin
44. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica 50:987–1008
45. Engle RF (2002) Dynamic conditional correlation – a simple class of multivariate garch models. J Bus Econ Stat 20: 339–350
46. Engle RF (2002) New frontiers of ARCH models. J Appl Econom 17:425–446
47. Engle RF, Bollerslev TP (1986) Modelling the persistence of conditional variances. Econom Rev 5:1–50, 81–87
48. Engle RF, Gonzalez-Rivera G (1991) Semiparametric ARCH models. J Bus Econ Stat 9:345–360
49. Engle RF, Kroner KF (1995) Multivariate simultaneous generalized arch. Econom Theory 11:122–150
50. Engle RF, Mezrich J (1996) GARCH for groups. RISK 9:36–40
51. Engle RF, Ng VK (1993) Measuring and testing the impact of news on volatility. J Finance 48:1749–1778
52. Engle RF, Lilien DM, Robins RP (1987) Estimating time varying risk premia in the term structure: The ARCH-M model. Econometrica 55:391–407
53. Engle RF, Ng VK, Rothschild M (1990) Asset pricing with a factor-ARCH covariance structure. J Econom 45:213–237
54. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–105
55. Fan J, Wang M, Yao Q (2008) Modelling multivariate volatilities via conditionally uncorrelated components. J Royal Stat Soc Ser B 70:679–702
56. Fiorentini G, Sentana E, Shephard N (2004) Likelihood-based

estimation of latent generalized ARCH structures. Econometrica 72:1481–1517

57. Francq C, Zakoian JM (2004) Maximum likelihood estimation of pure garch and arma-garch processes. Bernoulli 10:605–637
58. Fridman M, Harris L (1998) A maximum likelihood approach for non-gaussian stochastic volatility models. J Bus Econ Stat 16:284–291
59. Glosten LR, Jagannathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. J Finance 48:1779–1801
60. Gouriéroux C (1992) Modèles ARCH et Applications Financières. Economica
61. Gouriéroux C, Monfort A (1992) Qualitative threshold ARCH models. J Econom 52:159–199
62. Granger CWJ (1980) Long memory relationships and the aggregation of dynamic models. J Econom 14:227–238
63. Hafner CM (1998) Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models. J Stat Plan Inference 68:247–269
64. Hafner CM (2008) Temporal aggregation of multivariate GARCH processes. J Econom 142:467–483
65. Hafner CM, Rombouts JVK (2007) Semiparametric multivariate volatility models. Econom Theory 23:251–280
66. Hall P, Yao P (2003) Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica 71:285–317
67. Hansen PR, Lunde A (2006) Realized variance and market microstructure noise, with comments and rejoinder. J Bus Econ Stat 24:127–218
68. Härdle W, Tsybakov A (1997) Local polynomial estimation of the volatility function. J Econom 81:223–242
69. Harvey AC, Ruiz E, Shepard N (1994) Multivariate stochastic variance models. Rev Econ Stud 61:247–264
70. He C, Teräsvirta T (1999) Statistical Properties of the Asymmetric Power ARCH Process. In: Engle RF, White H (eds) Cointegration, Causality, and Forecasting. Festschrift in honour of Clive W.J. Granger, Oxford University Press, pp 462–474
71. Hentschel L (1995) All in the family: Nesting symmetric and asymmetric garch models. J Financial Econ 39:71104
72. Hill B (1975) A simple general approach to inference about the tail of a distribution. Ann Stat 3:1163–1174
73. Hong Y, Tu J, Zhou G (2007) Asymmetries in stock returns: Statistical tests and economic evaluation. Rev Financial Stud 20:1547–1581
74. Hull J, White A (1987) The pricing of options on assets with stochastic volatilities. J Finance 42:281–300
75. Jacquier E, Polson NG, Rossi PE (1994) Bayesian analysis of stochastic volatility models (with discussion). J Bus Econ Stat 12:371–417
76. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J Econom 122:185–212
77. Jeantheau T (1998) Strong consistency of estimators for multivariate arch models. Econom Theory 14:70–86
78. Jorion P (2000) Value-at-Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York
79. Kawakatsu H (2006) Matrix exponential GARCH. J Econom 134:95–128
80. Kim S, Shephard N, Chib S (1998) Stocahstic volatility: likelihood inference and comparison with ARCH models. Rev Econ Stud 65:361–393
81. King M, Sentana E, Wadhwani S (1994) Volatility and links between national stock markets. Econometrica 62:901–933
82. Kristensen D, Linton O (2006) A closed-form estimator for the garch(1,1) model. Econom Theory 323–337
83. Lanne M, Saikkonen P (2007) A multivariate generalized orthogonal factor GARCH model. J Bus Econ Stat 25:61–75
84. Ledoit O, Santa-Clara P, Wolf M (2003) Flexible multivariate GARCH modeling with an application to international stock markets. Rev Econs Stat 85:735–747
85. Lee SW, Hansen BE (1994) Asymptotic properties of the maximum likelihood estimator and test of the stability of parameters of the GARCH and IGARCH models. Econom Theory 10:29–52
86. Ling S, McAleer M (2003) Asymptotic theory for a vector ARMA-GARCH model. Econom Theory 19:280–310
87. Lintner J (1965) Security prices, risk and maximal gains from diversification. J Finance 20:587–615
88. Linton O (1993) Adaptive estimation in ARCH models. Econom Theory 9:539–569
89. Linton O, Mammen E (2005) Estimating semiparametric ARCH models by kernel smoothing methods. Econometrica 73:771–836
90. Lo A, Wang J (1995) Implementing option pricing models when asset returns are predictable. J Finance 50:87–129
91. Lumsdaine RL (1996) Asymptotic properties of the quasi maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models. Econometrica 64:575–596
92. Mahieu R, Schotman P (1998) An empirical application of stochastic volatility models. J Appl Econom 13:333–360
93. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36:394–419
94. Meddahi N, Renault E (2004) Temporal aggregation of volatility models. J of Econom 119:355–379
95. Morgan JP (1996) Riskmetrics Technical Document, 4th edn. J.P. Morgan, New York
96. Morillo D, Pohlman L (2002) Large scale multivariate GARCH risk modelling for long-horizon international equity portfolios. Proceedings of the 2002 Forecasting Financial Markets conference, London
97. Nelson DB (1990) ARCH models as diffusion approximations. J Econom 45:7–38
98. Nelson DB (1990) Stationarity and persistence in the GARCH(1,1) model. Econom Theory 6:318–334
99. Nelson DB (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
100. Nelson DB, Cao CQ (1992) Inequality constraints in the univariate garch model. J Bus Econ Stat 10:229–235
101. Newey WK, Steigerwald DS (1997) Asymptotic bias for quasi maximum likelihood estimators in conditional heteroskedasticity models. Econometrica 3:587–599
102. Nijman T, Sentana E (1996) Marginalization and contemporaneous aggregation in multivariate GARCH processes. J Econom 71:71–87
103. Pantula SG (1988) Estimation of autoregressive models with ARCH errors. Sankhya Indian J Stat B 50:119–138
104. Peng L, Yao Q (2003) Least absolute deviations estimation for ARCH and GARCH models. Biometrika 90:967–975
105. Robinson PM (1991) Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. J Econom 47:67–84

106. Ross SA (1976) The arbitrage theory of capital asset pricing. J Econ Theory 13:341–360
107. Sentana E, Fiorentini G (2001) Identification, estimation and testing of conditionally heteroskedastic factor models. J Econom 102:143–164
108. Sharpe WF (1964) Capital asset prices: A theory of market equilibrium under conditions of risk. J Finance 19:425–442
109. Shephard N (2005) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
110. Straumann D, Mikosch T (2006) Quasi-mle in heteroscedastic times series: a stochastic recurrence equations approach. Ann Stat 34:2449–2495
111. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
112. Tse YK (2000) A test for constant correlations in a multivariate GARCH model. J Econom 98:107–127
113. Tse YK, Tsui AKC (2002) A multivariate GARCH model with time-varying correlations. J Bus Econ Stat 20:351–362
114. van der Weide R (2002) Go-garch: A multivariate generalized orthogonal GARCH model. J Appl Econom 17:549–564
115. Vrontos ID, Dellaportas P, Politis DN (2000) Full bayesian inference for GARCH and EGARCH models. J Bus Econ Stat 18:187198
116. Vrontos ID, Dellaportas P, Politis D (2003) A full-factor multivariate garch model. Econom J 6:311–333
117. Wang Y (2002) Asymptotic nonequivalence of GARCH models and diffusions. Ann Stat 30:754–783
118. Weiss AA (1986) Asymptotic theory for ARCH models: Estimation and testing. Econom Theory 2:107–131
119. Yang L (2006) A semiparametric GARCH model for foreign exchange volatility. J Econom 130:365–384
120. Yang L, Härdle W, Nielsen P (1999) Nonparametric autoregression with multiplicative volatility and additive mean. J Time Ser Anal 20:579–604
121. Zaffaroni P (2007) Aggregation and memory of models of changing volatility. J Econom 136:237–249
122. Zaffaroni P (2007) Contemporaneous aggregation of GARCH processes. J Time Series Anal 28:521–544
123. Zakoian JM (1994) Threshold heteroskedastic functions. J Econ Dyn Control 18:931–955

Books and Reviews

Andersen T, Bollerslev T, Diebold F (2004) Parametric and nonparametric measurement of volatility. In: Ait-Sahalia L, Hansen LP (eds) Handbook of Financial Economtrics. Amsterdam (forthcoming)

Bauwens L, Laurent S, Rombouts J (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109

Bera A, Higgins M (1993) A survey of ARCH models: properties, estimation and testing. J Econ Surv 7:305–366

Bollerslev T, Chou R, Kroner K (1992) ARCH modelling in finance: a review of the theory and empirical evidence. J Econom 52:5–59

Bollerslev T, Engle R, Nelson D (1994) ARCH models. In: Engle R, McFadden D (eds) Handbook of Econometrics. North Holland Press, Amsterdam, pp 2959–3038

Engle R (1995) ARCH: Selected Readings. Oxford University Press, Oxford

Gouriéroux C (1997) ARCH Models and Financial Applications. Springer, New York

Shephard N (1996) Statistical aspects of ARCH and stochastic volatility. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London, pp 1–67

Shephard N (2005) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford

Taylor S (1986) Modelling Financial Time Series. Wiley, Chichester

Group Model Building

Etiënne A. J. A. Rouwette, Jac A. M. Vennix
Institute for Management Research, Radboud University,
Nijmegen, The Netherlands

Article Outline

Glossary

Facilitator Person who guides the group process in group model building.

Gatekeeper Person who forms the linking pin between modeling team and management team.

Knowledge elicitation Process of capturing the knowledge contained in the mental models of team members of the management team.

Modeler Person who constructs the quantified model during group model building.

Recorder Person who takes notes during group model building sessions and constructs workbooks.

Reference mode Graph(s) showing the behavior of the problem over time.

Workbook Booklet which contains summary of previous group model building sessions and prepares for subsequent sessions.

Client Person (or agency) who buys a model.

Definition of the Subject

Computer (simulation) models have been used to support policy and decision making in the decades after World War II. Over the years modelers learned that the application of these models to policy problems was not as straightforward as had been thought initially. As of the beginning of the 1970s studies started to appear that questioned the use of large-scale computer models to support policy and decision making (cf. [24,31]). Lee's article bears the significant title: ”Requiem for large scale models“, a statement that leaves little room for ambiguity. Other authors who have studied the impact records of computer models also seem rather sceptical (e. g. [9,22,25,73]). It is interesting to note that Greenberger et al., after interviewing both modelers and policy makers (for whom the models were constructed) found that modelers generally pointed to the fact that they learned a lot from modeling a particular policy issue. Policy makers on the other hand indicated that they did not really understand the models nor had much confidence in them. The results of these studies pointed in the direction of learning from computer models, i. e. conceptual or enlightenment use rather than instrumental use, where policy recommendations could straightforwardly be deduced from the model analysis and outcomes. In other words it is in the process of modeling a policy problem where the learning takes place which is required to (re)solve a problem. And it is also in this process that one needs to anticipate the implementation of policy changes. By the end of the 1970s system dynamics modelers pointed out that implementation of model outcomes was a neglected area (e. g. [50,74]) and that modelers sometimes naively assumed that implementation was straightforward, thereby neglecting organizational decision making as a political arena.

In other words client participation in the process of model construction and analysis is required for successful modeling and implementation of insights from the model into policy making. Or as Meadows and Robinson put it:

> Experienced consultants state that the most important guarantee of modelling success is the interested participation of the client in the modelling process (p. 408 in [34]).

Over the years this has given rise to all kinds of experiments to involve clients in the process of model construction. In the 1990s the term Group Model Building was introduced to refer to more or less structured approaches for client involvement in system dynamics model construction and analysis.

Introduction

From the early days of the field, the topic of client involvement in the process of model construction has raised attention in the system dynamics literature. Jay Forrester, the founder of the field of system dynamics, has repeatedly indicated that most of the knowledge needed to construct a system dynamics models can be found in the mental database of the participants of the system to be modeled [20,21]. Over the years several system dynamics modelers have experimented with approaches to involve client (groups) in model construction and analysis. This development in the system dynamics community parallels a movement in the operational research and systems fields towards more attention for stakeholders' opinions. A number of authors (e. g. [1]) criticized traditional OR and systems approaches as unsuitable for ill-structured problems that arise from differences between stakeholders' views on the problem. For ill-structured problems a range of new methods was developed [35].

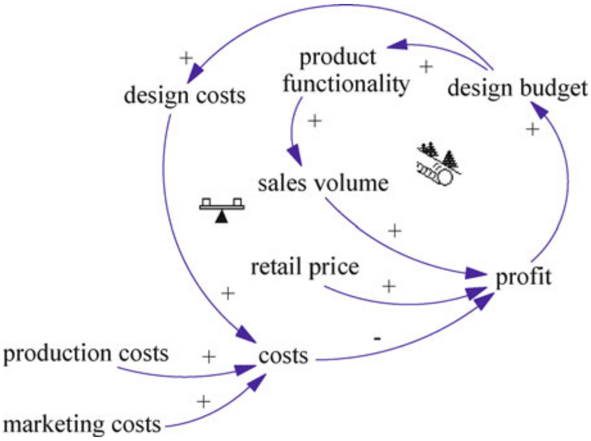
The developments in the system dynamics, operational research and systems communities have given rise to a set of distinct methods and approaches. However, practitioners work on problems that have clear similarities to those encountered in other disciplines and frequently borrow techniques from one another. The boundaries between methods are therefore difficult to draw and there is a degree of overlap between approaches in and between fields. Below we first describe the distinguishing characteristics of system dynamics, as this separates group model building most clearly from other approaches fostering client involvement. We then describe a number of distinct group model building approaches.

System Dynamics

System dynamics is most easily characterized by its emphasis on two ideas: (a) the importance of closed loops of information and action for social systems, i. e. social systems as information feedback systems and (b) the need to use formal models to study these loops. System dynamicists assume that the dynamic behavior of a social system is the result of its underlying feedback structure. Actors use the information about the structure as input to their decisions, and by implementing their decision influence system behavior. This creates an interlocked chain of action and information which is also known as a feedback loop. Richardson (see p. 1 in [44]) describes a feedback loop as follows:

> The essence of the concept … is a circle of interactions, a closed loop of action and information. The patterns of behavior of any two variables in such a closed loop are linked, each influencing, and in turn responding to the behavior of the other.

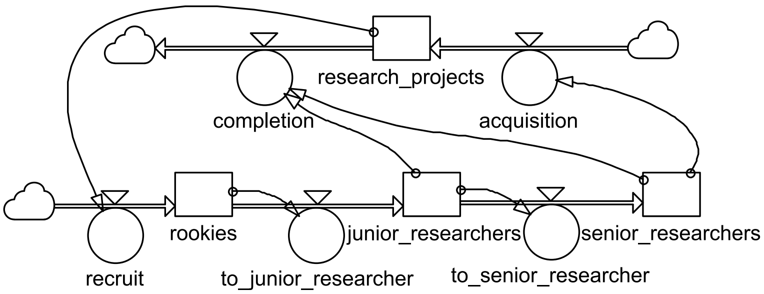
As an illustration of the use of information on the system state in decisions, imagine a simple example on customer behavior. Let us assume that if customers perceive that a product's functionality increases, more products will be bought. This will increase profits and thereby the design budget. An increased design budget can be used to improve the product's design, which will lead more customers to buy the product, and so on. Thus, decisions of actors within the system have an important influence on the system's behavior. If we continue to add other factors and relations to our example and capture these in a model, the diagram in Fig. 1 may result.



Group Model Building, Figure 1
Example of a causal loop diagram

As Fig. 1 shows, a causal loop diagram consists of variables, relationships, and feedback loops. Relations can be of two types: positive and negative. A positive relation indicates that both variables change in the same direction. In the model above, an increase in retail price will lead to an increase in profits, indicating a positive relationship. Variables in a negative relationship change in opposite directions. An increase in costs will decrease profits, indicating a negative relationship. The snowball rolling down the slope in the right hand side of Fig. 1 indicates a positive feedback loop. We assumed that an increase in profit results in a direct increase in the design budget. A higher budget allows for increased product functionality, which increases sales volume and finally profit. Starting from an increase in profit, the result is a further increase in profit. This is a so-called positive or self-reinforcing loop. However, if we assume that the design department uses its complete budget each year, an increased budget will contribute to design costs and lower profits. This is a negative or balancing loop, indicated by the balance symbol.

The second important idea in system dynamics is that formal models are necessary to understand the consequences of system structure. Since system dynamics models contain many (often non-linear) relations and feedback loops, it becomes very difficult to predict their behavior without mathematical simulation. Systems are assumed to consist of interacting feedback loops, which may change in dominance over time. Diagrams such as the one depicted above are frequently used in the interaction with clients. Before the dynamic consequences of the structure captured in Fig. 1 can be studied, it is necessary to further specify both the variables and relations used in the model.



Group Model Building, Figure 2
Example of a stock and flows diagram

Two categories of variables are distinguished: stocks and flows. Stocks are entities existing at a certain time period, for example supplies, personnel, or water in a reservoir. Flows are entities measured over a time period, such as deliveries, recruitment, or inflow of water. Relationships are separated into physical flows and information flows. If we capture differences between stocks and flows and information and physical flows in a diagram, a stock and flows diagram results.

As can be seen in Fig. 2, information links are depicted with a single arrow and physical flows with a double arrow. The physical human resources flow is separated in three stocks: number of rookies, number of junior researchers, and number of senior researchers. Recruitment will lead to an increase in the number of rookies. Two other flows influence the number of people in the stocks: rookies may be promoted to junior researchers and junior researchers may be promoted to senior researchers. The human resources flow is related to the project flow with information links, for instance indicating that acquisition of research projects is determined by the number of senior researchers.

Group Model Building Approaches

As pointed out before, client involvement has been important to system dynamics from the start of the field. System dynamics emphasizes feedback loops and the use of formal models. In this section we describe how models based on these ideas are built in interaction with actors and stakeholders in the problem at hand. A number of different participative model building formats can be identified, i. e. the reference group approach [43,61]; the stepwise approach [76]; the Strategic Forum [49]; modeling as learning [29]; strategy dynamics [71,72] and Hines' "standard method" [39]. Below we describe each approach briefly.

In the Reference Group approach [43,61] participation takes the form of frequent interaction between the modeling team and a group of eight to ten clients. The approach starts with the identification of interest groups, of which representatives are invited to contribute to the modeling effort. The representatives are referred to as referents. In a series of interviews and meetings, the problem to be addressed is defined more specifically. On the basis of this definition and the information gathered in the interviews and meetings, the modeling team develops a preliminary model. In the remainder of the project the modelers are responsible for model improvements while the referents function as critics. This model is elaborated in a series of meetings and is at the same time used as a tool for structuring the discussion. In later sessions, model output is used for developing scenarios. In a scenario discussion the model is run and results are described and analyzed by the modelers. The reference group is then asked to determine to what extent the model's behavior corresponds to their expectations about reality, and if it does not, to suggest changes. These suggestions can trigger changes in the model structure, initiating a new round in the discussions.

The stepwise approach [76] is founded on the idea that full quantification of models is not always possible or desirable. The approach starts with a definition of the problematic behavior. If possible, this definition is given in the form of a behavior over time of the problem of interest. Modeling starts by roughly sketching the feedback loops responsible for this behavior. The key variables related to the cause for concern are identified, followed by the system resources connected to these key variables and their initial states. The resources are used to derive the central stocks in the system. From the resources, the resource flows can then be sketched with the associated rates of conversion. Delays are added to these flows if they are significant. Next, organizational boundaries, flows of information and strategies through which the stocks influence the flows, are added. Again, if there are significant delays, these are added to the information linkages. In the final step, information flows and strategies linking different resource flows are added. The steps are repeated until the relevant feedback loops have all been included. Wolstenholme indicates that these steps often provide the insights necessary to infer system behavior from the structure, which reduces the need for quantification. Models can also be analyzed in a qualitative manner.

The steps that make up the Strategic Forum [49] provide a detailed insight of how clients are encouraged to participate in modeling. The strategic forum consists of eight steps, of which the first two are conducted before the actual meeting (also called the forum) with the client group. The process begins with interviews prepared by a small questionnaire, in which three issues are addressed: ideas on the current situation, a statement of the vision for the future, and agreement on a preliminary map of the problem. On the basis of the interviews, the modeler constructs an integrated map and accompanying computer model. In the second step the project team designs a number of small group exercises that will be used during the forum. The exercises are aimed at discovering important structural and behavioral elements and are similar to the scenario discussions in the reference group approach. The most important difference is that before simulation results are shown, participants have to 'put a stake in the ground', i. e. they have to make a prediction of model behavior on the basis of a change in a policy variable and values for connected parameters. The model is then simulated and results are compared with participants' expectations. Discrepancies between predictions and simulations are identified, and might point to inconsistencies in participants' ideas or lead to model improvements. In the following steps the participants meet in a series of workshops. Each workshop opens with an introduction and a big picture discussion. The heart of the session consists of exercises aimed at internal consistency checks, addressing the consistency between the group's mental model and the computer model. As in the other approaches, model structure will be changed if inconsistencies with the participants' ideas on the problem are revealed. In the final phase of policy design, potential consequences of strategic policies are addressed and the existing capability of realizing the strategic objectives. A wrap-up discussion and identification of follow-up activities concludes the Strategic Forum.

Richmond (see p. 146 in [49]) emphasizes that the main purpose of the Strategic Forum is to check the consistency of strategy. The insights gained by the client therefore frequently lead to changes in strategy or operating policies, but less frequently to changes in objectives or the mission statement. One important element of ensuring an impact on participants' ideas is the (dis)confirmation of expectations on simulation outcomes.

Lane [29] describes a modeling approach developed at Shell International Petroleum, known as 'modeling as learning'. Lane explicitly sets this approach apart from the widely used expert consultancy methodology (e. g. [58]). His approach also puts strong emphasis on involving decision makers in the modeling process. By showing decision makers the benefits of participation early on in the process, an attempt is made to persuade them to spend time in direct interaction with the model. The approach centers on capturing and expressing the client's ideas, initiating a discussion on the issue with 'no a-priori certainty regarding quantification, or even cause and effect' (see p. 70 in [29]). The modelers also strive to include both hard as well as 'soft' aspects of the problematic situation. In doing this, it is hoped that the clients' ideas are included in the model and that ownership is created. This is encouraged by making models and model output transparent to participants, helping the client 'to learn whichever techniques are used in a project' (see p. 71 in [29]). Lane states that the focus throughout the approach is on a process of learning, using such elements as experimentation with the model, testing of assumptions and representing and structuring ideas in a logical way.

Hines' approach [39] starts off by diagnosing the problem. This step comes down to gathering and clustering problem variables. Problem behavior is visualized by sketching the graph over time of the problematic behavior. In the second step the structure underlying the problematic behavior is captured in a causal diagram. This so-called dynamic hypothesis incorporates many of the problem variables identified earlier. The diagram helps to clarify the boundary of the problem that will be addressed and thus limits the project scope. The next step is to identify accumulations in the system, which will form the stocks in the system dynamics model. In the construction of the computer model most work is done by the modelers, with client participation limited to providing data such as numerical values and details of the work processes relevant to the problem at hand. Model structure and behavior is then explained to the client. Discussions with the client then lead to a series of model iterations, increasing confidence of the client in model calibration and validity. Similar to other participative approaches, policy runs are used to test proposed interventions in the problem.

Warren [71,72] describes an approach to participative modeling that strongly focuses on identifying accumulations (stocks) in the system. In order to identify central accumulations, clients are asked to identify the strategic resources in the problem at hand. Increases and decreases in resources then lead to the identification of flows. Warren's approach differs from the ones described above in the sense that stocks and flows are differentiated from the outset. This means that causal loop diagrams are not used. In addition, graphs over time are recorded next to each variable in the model. By gradually adding elements to the model while visually relating structure and behavior, the clients' understanding of the problem is gradually increased.

As mentioned before, the boundaries around approaches are not easy to draw and one method may 'borrow' techniques of another. Insights and practices from the operational research and system fields have been merged with those in system dynamics to develop combined methods. For example, modeling as learning is one of the approaches incorporating elements of soft operational research methodologies. Lane and Oliva [30] describe the theoretical basis for integrating system dynamics and soft systems methodology. The cognitive mapping approach (e. g. [17]) also offers tools and techniques that are used in system dynamics studies.

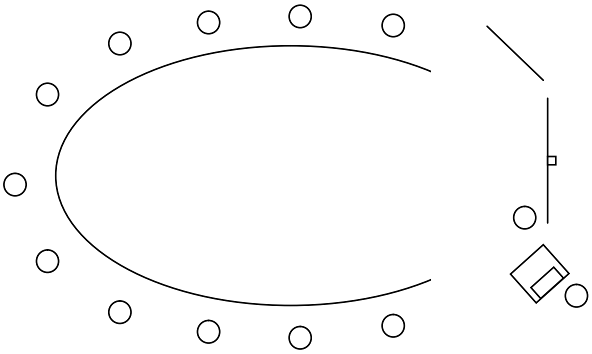
In addition to combining different methods, approaches are sometimes also tailored to use in specific content areas. An example is van den Belt's [63] mediated modeling, which combines insights from participative system dynamics modeling and consensus building on environmental issues.

Group Model Building: Basic Ideas and Concepts

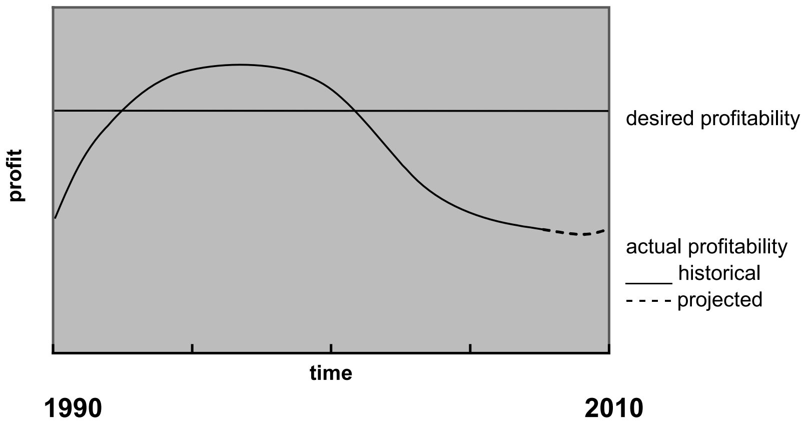
The separate approaches described in the last section continue to be developed and used in practical problems. Although we are not sure that all proponents of these approaches would characterize themselves as using "group model building", this term has been used more and more in the last decades to refer to system dynamics approaches with client involvement in a general sense. The two approaches that coined the term group model building evolved more or less simultaneously, with considerable cross-fertilization of ideas, at SUNY at Albany and Radboud University Nijmegen in the Netherlands (see [36,69]). In an early application at Radboud University, participants were involved in a Delphi study consisting of mailed questionnaires and workbooks, followed by workshops [66]. In the dissertations by Verburgh [70] and Akkermans [4] a similar approach is used under the name of participative policy modeling and participative business modeling, respectively. In its latest version group model building is a very open approach, which allows for the use of preliminary models or a start from scratch, uses individual interviews, documents and group sessions, qualitative or quantitative modeling and small as well as large models. Vennix [64,65] provides a set of guidelines for choosing among these different approaches, building on and adding to the studies mentioned above. Andersen and Richardson [5] provide a large number of "scripts" that can help in setting up modeling projects. The procedures described are a long way from the earlier descriptions of a set of steps that seem to prescribe standard approaches applicable to most modeling projects. Instead, the guidelines offered have more the appearance of tool boxes, from which the appropriate technique can be selected on the basis of problem characteristics and the clients involved.

Group model building is generally conducted with a group of at least six and up to 15 people. The group is guided by at least two persons: a facilitator and a modeler/recorder. The group is seated in a semi circle in front of a whiteboard and/or projection screen, which serves as a so-called group memory. A projection screen is typically used when a model is constructed with the aid of system dynamics modeling software with a graphic interface (e. g. Vensim, Powersim, Ithink). This group memory documents the model under construction and is used as a parking lot for all kinds of unresolved issues which surface during the deliberations of the group.

In Fig. 3, the small circles indicate the persons present in the session. Apart from the participants, there is a facilitator and a recorder. The facilitator has the most important role in the session as he or she guides the group process. His/her task, as a neutral outsider, is to (a) elicit relevant knowledge from the group members, (b) to (help) translate elicited knowledge into system dynamics modeling terms, and (c) make sure that there is an open communication climate so that in the end consensus and commitment will result. The recorder keeps track of the elements of the model. In Fig. 3 (s)he is seated behind a computer and the model is projected on the screen in front of the



Group Model Building, Figure 3
Typical room layout for group model building with participants seated in a semi-circle, white board and facilitator in front, and computer and overhead projector (adapted from [5])



Group Model Building, Figure 4
Problem description in graphical form: reference mode of behavior

group. A separate whiteboard (upper right hand corner) is used to depict the reference mode of behavior and record comments or preliminary model structure. As the model is visible to all participants, it serves as a group memory that at each moment reflects the content of the discussion up to that point. A group model building session is generally conducted in the so-called chauffeured style, where only the facilitator uses electronic support and projection equipment, while participants do not have access to electronic communication media [38]. The central screen or whiteboard will be used to depict the model, as shown in Fig. 3.

The role of liaison between the organization and the modeling team is performed by the gatekeeper, who is generally also a member of the participant group. The gatekeeper is the contact between both parties, and has an important role in the decision which participants to involve in the sessions. Apart from the gatekeeper, the facilitator and the recorder, two other roles may be important in a modeling session [46], i. e. a process and a modeling coach. The process coach functions as an observer and primarily pays attention to the group process. The model coach needs to be experienced in system dynamics modeling but might also be an expert in the content area as well. As Richardson and Andersen [46] point out, all roles are important in group model building but not all of them have to be taken up by a single person. One person might for instance combine the roles of facilitator and process coach. Taken together, these different roles constitute the facilitation or modeling team.

In principle the group follows the normal steps in the construction of a system dynamics model. This means that the first step is the identification of the strategic issue to be discussed, preferably in the form of a so-called reference mode of behavior, i. e. a time series derived form the system to be modeled which indicates a (historical) undesirable development over time. As an example let us take the sales of a software product. An initial problem statement might be falling profit. Typically the problematic behavior will be depicted in a graph over time as in Fig. 4.

In the graph above, a projection of the expected behavior is included for the years after 2008.

The next step is to elicit relevant variables with which the model construction process can be started. Depending on the type of problem this will take the form of either a causal loop diagram or a stocks and flow diagram and is generally referred to as the conceptualization stage. The following step is to write mathematical equations (model formulation) and to quantify the model parameters. As described in the introductory section, most of the model formulation work is done backstage as it is quite time consuming and members of a management team generally are not very much interested in this stage of model construction. In this stage, the group is only consulted for crucial model formulations and parameter estimations. Experienced group model builders will start to construct a simple running model as soon as possible and complicate it from there on if required. In the end the model should of course be able to replicate the reference mode of behavior (as one of the many validity tests) before it can be sensibly be used as a means to simulate the potential effects of strategies and scenarios.

Objectives of Group Model Building

As mentioned in the introduction, the founder of system dynamics has repeatedly pointed out that much of the knowledge and information which is needed to construct a model can be found in the mental models of system participants. At first sight it may seem that the most important objective of building a system dynamics model is to find a robust strategy to solve the problem of the organization. In the end that is why one builds these models. From this perspective the most important issue in group model building is how to elicit the relevant knowledge from the group. However, as stated before, decision making in organizations has its own logic, and in many cases there is quite some disagreement about the problem and how it should be tackled. No wonder that implementation of model outcomes is difficult if the model building process is not well integrated with decision making processes in organizations, when it comes to creating agreement and commitment with a decision. From that perspective knowledge elicitation is only one element in the process of model construction. It is not so much the model but to a greater extent the process of model construction which becomes important. Somewhat simplified one could say that in the "standard" approach when an organization is confronted with a strategic problem it hires a modeler (or group of modelers) to construct a model and come up with recommendations to "solve" the problem. However, in most cases these recommendations become part of the discussion in the management team and get misunderstood, or adapted or frequently just disappear as a potential solution from the discussion. Hence Watt's title of his paper: "Why won't anyone believe us?" becomes very much understandable from the point of view of the modeler. So rather than creating a situation where modelers "take away" the problem from the organization and (after considerable time) return with their recommendations, the model building process is now used to structure the problem, guide communication about it and test the robustness of strategies taking into account other criteria and information which is not included in the model, but does play a role for the organization when making the decision. Stated differently, the model building process now becomes intertwined with the process of decision making in an organization. And this in turn means that other objectives than knowledge elicitation become important.

Simultaneously with the attempts to involve clients in the process of constructing system dynamics models the objectives of group model building have been defined at several levels, i. e. the individual, the group and the organizational level (cf. [7,65]). The main goal at the individual level is change of mental models and learning. The idea is that participants should better understand the relationship between structure and dynamics and how their interventions may create counterintuitive results. Unfortunately research has revealed that this is hardly the case. Even after extensive training people have difficulty to understand the relationship between structure and dynamics (for a review see [8,53]). A second goal at the individual level is behavioral change. Frequently the conclusions of a modeling intervention point in the direction of behavioral change, for example implementing a new job rotation scheme, or a change in purchasing policy. The question can then be asked how insights from the modeling intervention are translated to changes in behavior. Rouwette (e. g. [52,53,57]) uses a framework from social psychology to understand the impact of modeling on behavior. The theory of Ajzen [2,3] explains behavior on the basis of (a) attitude, (b) perceptions of norms and (c) perceptions of control. It seems likely that each of these concepts is influenced in modeling sessions. When for example model simulations reveal unexpected levers for improving system behavior, we can expect that perceived control will increase. Another example: let's imagine that a manager is participating in a modeling session, where another participant reveals positive outcomes of a certain policy option. If these positive outcomes were previously not known to the manager, hearing them might make his/her attitude towards that option more positive (cf. [42]).

At the group level objectives refer to mental model alignment [28] and fostering consensus [51,67,75]). Creating consensus should not be confused with premature consensus, i. e. not discussing conflicting viewpoint. Here it concerns creation of consensus after critical debate and discussion of opinions has taken place. This type of discussion which needs to take place in a cooperative communication climate is helpful to also create commitment with the resulting decision.

At the organizational level goals have been discussed as system process change (are things done differently) and system outcome change (are customers impacted differently) [11]. Although it has to be pointed out that in many cases system changes are the result of changes in attitude and behavior of participants in the system. An overview of group model building objectives is given in Table 1. In this table finds a number of additional objectives such as positive reaction and creation of a shared language, that are more fully reviewed by Huz et al. [28], Rouwette et al. [55] and Rouwette and Vennix [53].

Designing Group Model Building Projects

When designing group model building projects there are a number of questions that need to be addressed. The first concerns the suitability of system dynamics for the problem at hand. System dynamicists generally say that a prob-

Group Model Building, Table 1
Objectives of group model building (cf. [53])

Individual	Positive reaction
	Mental model refinement
	Commitment
	Behavioral change
Group	Increased quality of communication
	Creation of a shared language
	Consensus and alignment
Organization	System changes
	System improvement or results
Method	Further use
	Efficiency

lem needs to be dynamically complex in order to be suitable to model it through system dynamics. This means that one should at least hypothesize that there are positive and negative feedback process underlying the problem. From a more practical point of view one could say that one should be able to represent the problem in the form of a reference mode of behavior. If the latter is not possible one should seriously question the use of system dynamics for the problem.

A second issue which needs to be given some thought is the question whether to use qualitative or quantitative modeling. Within the system dynamics community there is still a debate about the question whether qualitative modeling (or: mapping) can be considered system dynamics (see [13,14,23]). In short those who disagree point out that without quantification and simulation one cannot reliably develop a robust policy simply because the human mind is not capable of predicting the dynamic effects of (interventions in) a dynamically complex structure. Those who do use mapping on the other hand point out that mapping in itself can have the beneficial effect to structure the problem and at least will make managers aware of potential underlying feedback loops and their potential counterintuitive effects when intervening in a dynamically complex system. Basically the issue to quantify or not depends on the goals of the group model building intervention. If the ultimate goal is to find robust policies then quantification is required. However, if the aim is to structure a problem and to create consensus on a strategic issue then qualitative modeling may be all that is needed. This links up with Zagonel's [77] distinction between the use of models as micro worlds or as boundary objects. When used as a boundary object the emphasis is on supporting negotiation and exchange of viewpoints in a group. This is clearly the case when problems are messy, i. e. connected to other problems and when there is much divergence of opinion on what the problem is and sometimes even whether there is a problem at all.

A third issue is the question who to involve in the sessions. There are a number of criteria which are generally employed. First it is important to involve people who have the power to make decisions and changes. A second criterion is to involve people who are knowledgeable about the problem at hand. A third criterion is to involve a wide variety of viewpoints, in order to make sure that all relevant knowledge about the problem is included. Of course these guidelines may create dilemmas. For example, involving more people in the process will make the group communication process more difficult. This may in turn endanger the creation of consensus and commitment.

Another issue is whether to use a preliminary model or to start from scratch (see [45]). Although using a preliminary model may speed up the process the inherent danger is that it will be difficult to build group ownership over the model. Group ownership is clearly required to create consensus and commitment.

Finally, a range of methods and techniques is available to elicit relevant knowledge both from individuals and from groups. When it comes to individuals, well known methods are interviews, questionnaires and so-called workbooks. The latter are a kind of modified questionnaires, which are used in between sessions to report back to the group and ask new question in preparation of the next session. Interviews are being used routinely as a preparation for group model building sessions.

If a decision is made on the issues discussed above, the next important question is how to plan and execute the modeling sessions. This question is a central topic in the group model building literature and its success heavily depends on the correct choice of available techniques and the quality of the facilitator.

Conducting Group Model Building Sessions

Although careful preparation of group model building sessions is a necessity, the most important part of the whole project is what happens in the group model building sessions themselves. During the sessions not only the analysis of the problem takes place (and the model is constructed), but also the interaction process between members of the management team unfolds. It is this interaction process which needs to be guided in such a way that consensus and commitment will emerge and implementation of results will follow. As pointed out the process is guided by the group facilitator, generally someone who is not only specialized in facilitation of group processes but also in system dynamics model construction. The facilitator is supported

Group Model Building, Table 2
Group model building scripts (cf. [5])

Phase in modeling	Script
Defining a problem	Presenting reference modes
	Eliciting reference modes
	Audience, purpose, and policy options
Conceptualizing model structure	Sectors, a top down approach
	Maintain sector overview while working within a sector
	Stocks and flows, by sector
	Name that variable or sector
Eliciting feedback structure	Direct feedback loop elicitation
	Capacity utilization script
	System archetype templates
	"Black box" means-ends script
Equation writing and parametrization	Data estimation script
	Model refinement script
	"Parking lot" for unclear terms
Policy development	Eliciting mental model-based policy stories
	Create a matrix that links policy levers to key system flows
	"Complete the graph" policy script
	Modeler/reflector feedback about policy implications
	Formal policy evaluation using multiattribute utility models
	Scripts for "ending with a bang"

by a recorder or modeler who helps constructing the system dynamics model while the facilitator interacts with the management team.

The facilitator may choose from a wide variety of techniques in setting up and conducting a session. As a foundation for choosing techniques, Andersen and Richardson [5] develop a set of guiding principles and so-called scripts for group model building sessions. Guiding principles capture basic ideas in the interaction with clients, such as break task/group structure several times each day, clarify group products, maintain visual consistency and avoid talking heads. Scripts are more concrete instances of these principles and refer to small elements of the interaction process [5,32]. The Table 2 shows scripts described in Andersen and Richardson's [5] original paper.

In choosing a script it is first important to be aware of the phase that is relevant in the project at that time. A common starting point, as we saw in the description of group model building approaches, is to define the central problem of interest. The reference mode of behavior can function as a guideline for involving clients in this phase. Once the central problem is clear, a logical next step is to move towards model conceptualization. In this step again a number of options are available. Andersen and Richardson [5] describe a script for identifying sectors that are important in the problem. An alternative is to start with more concrete variables in the problem, using a Nominal Group Technique [15].

Whatever scripts and techniques a facilitator employs it is important that (s)he displays the right attitude and uses the correct skills. Several different aspects of attitude are important. First of all the facilitator is not the person who thinks (s)he knows the best solution, but needs to be helpful in guiding the group to find a solution to the strategic problem the organization is faced with. Second, a facilitator should be neutral with respect to the problem that is being discussed. Being too knowledgeable about a particular problem area (e. g. strategic alliances) may thus be dangerous, because it creates the tendency to participate in the discussions. Rather than being an expert, having an inquiry attitude (i. e. asking questions rather than providing answers) is more helpful to the group. Finally, integrity and being authentic is important. Relying on tricks to guide the process will prove counterproductive, because people will look through them.

When it comes to skills, a thorough knowledge and experience in constructing system dynamics models is of course indispensable. Second, a facilitator needs to be knowledgeable about group process and have the skills to structure both the strategic problem as well as the group interaction process. For the latter, special group process techniques (e. g. brainstorming, Nominal Group Technique, Delphi) may be used, and knowledge about and skills in applying these techniques is of course a prerequisite for a successful group model building intervention. Finally, communication skills are important. Reflective listening is a skill which will help to prevent misunderstanding in communication, both between participants and the facilitator and between group members. For a more thorough discussion of these attitudes and skill in the context of group model building we refer to Vennix [64].

Researching Group Model Building Effectiveness

In the previous sections we described goals of group model building projects and principles and scripts for guiding the modeling process. In this section we consider the empirical evidence for a relation between modeling interventions and these intended outcomes. Empirical evidence can be gathered using a variety of research strategies, such as (field) experiments, surveys or (in-depth) case studies. According to the review of modeling studies by Rouwette et al. [55], the case study is the most frequently used design to study group model building interventions. We first report on the results found by these authors and then turn to other designs.

In the meta-analysis of Rouwette et al. [55], the majority of group model building studies uses a case study design and assesses outcomes in a qualitative manner. Data are collected using observation, and a minority of studies employs individual group interviews. Case reports may be biased towards successful projects and are frequently not complete. The outcomes of the modeling projects were scored along the dimensions depicted in Table 1 in the section on modeling goals. The findings show positive outcomes in almost all dimensions of outcomes. Learning about the problem seems to be a robust outcome of group model building, for example:

- Of 101 studies that report on learning effects, 96 indicate a positive effect;
- Of 84 studies focusing on implementation of results, 42 report a positive effect.

Another set of studies, using quantitative assessment of results is described by Rouwette and Vennix [53]. Although the research surveyed so far indicates positive effects of modeling on outcomes such as mental model refinement, consensus and implementation of results, important challenges remain. Research so far has paid little attention to the complexity of the intervention as described in the previous section. Pawson and Tilley [40] urge us not to assume that interventions are similar and lead to similar effects, since this would confuse meaningful differences between studies. Rouwette and Vennix [53] describe two ways to learn more about the process through which outcomes of modeling are created: base research more on theory and/or to conduct research in more controlled settings. At present only few studies address elements of group model building in a controlled setting. Shields [59,60] investigates the effect of type of modeling and facilitation on a group task. Most research on the use of system dynamics models concerns so-called management flight simulators. These studies aim to mimic the important characteristics of decision making in complex, dynamic problems, and test the effectiveness of various decision aids. Results are reviewed by Sterman [62], Hsiao and Richardson [26] and Rouwette, Größler and Vennix [56].

Increased attention to theories may shed more light on the way in which modeling effects group decisions. Theories can help in specifying relations which can then be tested. Explanatory research is needed to connect the components and outcomes of group model building interventions (see p. 194 in [7]). In the field of system dynamics modeling, two attempts at formulating theories on modeling components and outcomes are the work of Richardson et al. [47] and Rouwette [52]. The framework formulated by Rouwette [57] builds on the theories of Ajzen [2,3] and Petty and Cacioppa [42] described earlier. Richardson et al. [47] separate mental models into means, ends and means-ends models. The ends model contains goals, while the means model consists of strategies, tactics, and policy levers. The means-ends model contains the connection between the two former types of models and may contain either detailed "design" logic or more simple "operator" logic. On the basis of research on participants in a management flight simulator [6], the authors conclude that operator logic, or high level heuristics, is a necessary condition for improving system performance. Therefore, providing managers with operator knowledge is the key to implementation of system changes.

Future Directions

The success of group model building and problem structuring methodologies in general depends on a structured interaction between theory, methodology refinement and application in practical project accompanied by systematic empirical evaluation.

Rouwette and Vennix [53] indicate three areas for further development of theories:

- Review related methodologies used in complex organizational problems, to determine which theories are used to explain effects. Examples that come to mind are theories used in the operational research and systems fields [37].

- Forge a closer connection to research on electronic meeting systems. In this field, studies are usually conducted in controlled settings [16,18,41] and theory development seems to be at a more advanced stage. Research on electronic meeting systems is interesting both because of the empirical results and explanatory theories used and because of insights on the intervention process. A recent development in the field is research on ThinkLets [10]. A ThinkLet is defined as a named, packaged facilitation intervention and thus seems very similar to the concept of a group model building script.
- A third source of theories is formed by research in psychology and group decision making. Theories from these fields inform the definition of central concepts in group model building (see Table 1) and theories on modeling effectiveness. Rouwette and Vennix [54] review literature on group information processing and relate this to elements of group model building interventions.

From theories and evaluation research will come insights to further develop the methodology along the lines of (a) determining what kind of problem structuring methodology is best suited in what kind of situation, (b) refinement of procedures, (c) better understanding the nature of the intervention, and (d) better guidelines for facilitators how to work in different kinds of groups and situations.

Bibliography

Primary Literature

1. Ackoff RA (1979) The future of operational research is past. J Oper Res Soc 30(2):93–104
2. Ajzen I (1991) The theory of planned behavior. Organ Behav Human Decis Process 50:179–211
3. Ajzen I (2001) Nature and operation of attitudes. Annu Rev Psychol 52:27–58
4. Akkermans HA (1995) Modelling with managers: Participative business modelling for effective strategic decision-making. Ph D Dissertation, Technical University Eindhoven
5. Andersen DF, Richardson GP (1997) Scripts for group model-building. Syst Dyn Rev 13(2):107–129
6. Andersen DF, Maxwell TA, Richardson GP, Stewart TR (1994) Mental models and dynamic decision making in a simulation of welfare reform. Proceedings of the 1994 International System Dynamics Conference: Social and Public Policy. Sterling, Scotland, pp 11–18
7. Andersen DF, Richardson GP, Vennix JAM (1997) Group model-building: Adding more science to the craft. Syst Dyn Rev 13(2):187–201
8. Andersen DF, Vennix JAM, Richardson GP, Rouwette EAJA (2007) Group model building: Problem structuring, policy simulation and decision support. J Oper Res Soc 58(5):691–694
9. Brewer GD (1973) Politicians, bureaucrats and the consultant: A critique of urban problem solving. Basic Books, New York
10. Briggs RO, de Vreede GJ, Nunamaker JFJ (2003) Collaboration engineering with ThinkLets to pursue sustained success with group support systems. J Manag Inf Syst 19:31–63
11. Cavaleri S, Sterman JD (1997) Towards evaluation of systems thinking interventions: A case study. Syst Dyn Rev 13(2): 171–186
12. Checkland P (1981) Systems thinking, systems practice. Wiley, Chichester/New York
13. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: Some research questions. Syst Dyn Rev 16(3):225–244
14. Coyle G (2001) Maps and models in system dynamics: Rejoinder to Homer and Oliva. Syst Dyn Rev 17(4):357–363
15. Delbecq AL, van de Ven AH, Gustafson DH (1975) Group techniques for program planning: A guide to nominal group and delphi processes. Scott, Foresman and Co, Glenview
16. Dennis AR, Wixom BH, van den Berg RJ (2001) Understanding fit and appropriation effects in group support systems via meta-analysis. Manag Inf Syst Q 25:167–183
17. Eden C, Ackermann F (2001) SODA – The principles. In: Rosenhead J, Mingers J (eds) Rational analysis for a problematic world revisited. Problem structuring methods for complexity, uncertainty and conflict. Wiley, Chichester, pp 21–42
18. Fjermestad J, Hiltz SR (1999) An assessment of group support systems experimental research: Methodology and results. J Manag Inf Syst 15:7–149
19. Fjermestad J, Hiltz SR (2001) A descriptive evaluation of group support systems case and field studies. J Manag Inf Syst 17:115–160
20. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
21. Forrester JW (1987) Lessons from system dynamics modelling. Syst Dyn Rev 3(2):136–149
22. Greenberger M, Crenson MA, Crissey BL (1976) Models in the policy process: Public decision making in the computer era. Russel Sage Foundation, New York
23. Homer J, Oliva R (2001) Maps and models in system dynamics: A response to Coyle. Syst Dyn Rev 17(4):347–355
24. Hoos IR (1972) Systems analysis in public policy: A critique. University of California Press, Berkeley/Los Angeles/London
25. House PW (1982) The art of public policy analysis: The arena of regulations and resources, 2nd printing. Sage, Thousand Oaks
26. Hsiao N, Richardson GP (1999) In search of theories of dynamic decision making: A literature review. In: Cavana RY et al (eds) Systems thinking for the next millennium – Proceedings of the 17th international conference of the system dynamics society. Wellington, New Zealand
27. Huz S (1999) Alignment from group model building for systems thinking: Measurement and evaluation from a public policy setting. Unpublished doctoral dissertation, SUNY, Albany
28. Huz S, Andersen DF, Richardson GP, Boothroyd R (1997) A framework for evaluating systems thinking interventions: an experimental approach to mental health system change. Syst Dyn Rev 13(2):149–169

29. Lane DC (1992) Modelling as learning: A consultancy methodology for enhancing learning in management teams. Eur J Oper Res 59(1):64–84. Special issue of Eur J Oper Res: Morecroft JDW, Sterman JD (eds) Modelling for learning
30. Lane DC, Oliva R (1998) The greater whole: Towards a synthesis of system dynamics and soft systems methodology. Eur J Oper Res 107(1):214–235
31. Lee DB (1973) Requiem for large-scale models. J Am Inst Plan 39(1):163–178
32. Luna-Reyes L, Martinez-Moyano I, Pardo T, Cresswell A, Andersen D, Richardson G (2006) Anatomy of a group model-building intervention: Building dynamic theory from case study research. Syst Dyn Rev 22(4):291–320
33. Maxwell T, Andersen DF, Richardson GP, Stewart TR (1994) Mental models and dynamic decision making in a simulation of welfare reform. In: Proceedings of the 1994 international system dynamics conference, Stirling, Scotland. Social and Public Policy. System Dynamics Society, Lincoln, pp 11–28
34. Meadows DH, Robinson JM (1985) The electronic oracle: Computer models and social decisions. Wiley, Chichester/New York
35. Mingers J, Rosenhead J (2004) Problem structuring methods in action. Eur J Oper Res 152:530–554
36. Morecroft JDW, Sterman JD (1992) Modelling for learning. Special issue Eur J Oper Res 59(1):28–41
37. Morton A, Ackermann F, Belton V (2003) Technology-driven and model-driven approaches to group decision support: focus, research philosophy, and key concepts. Eur J Inf Syst 12:110–126
38. Nunamaker JF, Dennis AR, Valacich JS, Vogel DR, George JF (1991) Electronic meetings to support group work. Commun ACM 34(7):40–61
39. Otto PA, J Struben (2004) Gloucester fishery: Insights from a group modeling intervention. Syst Dyn Rev 20(4):287–312
40. Pawson R, Tilley N (1997) Realistic evaluation. Sage, London
41. Pervan G, Lewis LF, Bajwa DS (2004) Adoption and use of electronic meeting systems in large Australian and New Zealand organizations. Group Decis Negot 13(5):403–414
42. Petty RE, Cacioppo JT (1986) The elaboration likelihood model of persuasion. Adv Exp Soc Psychol 19:123–205
43. Randers J (1977) The potential in simulation of macro-social processes, or how to be a useful builder of simulation models. Gruppen for Ressursstudier, Oslo
44. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
45. Richardson G (2006) Concept models. Größler A, Rouwette EAJA, Langer RS, Rowe JI, Yanni JM (eds) Proceedings of the 24th International Conference of the System Dynamics Society, Nijmegen
46. Richardson GP, Andersen DF (1995) Teamwork in group model-building. Syst Dyn Rev 11(2):113–137
47. Richardson GP, Andersen DF, Maxwell TA, Stewart TR (1994) Foundations of mental model research. Proceedings of the 1994 international system dynamics conference, Stirling, Scotland. pp 181–192. System Dynamics Society, Lincoln
48. Richmond B (1987) The strategic forum: From vision to strategy to operating policies and back again. High Performance Systems Inc, Highway, Lyme
49. Richmond B (1997) The strategic forum: Aligning objectives, strategy and process. Syst Dyn Rev 13(2):131–148
50. Roberts EB (1978) Strategies for effective implementation of complex corporate models. In: Roberts EB (ed) Managerial applications of system dynamics. Productivity Press, Cambridge, pp 77–85
51. Rohrbaugh JW (1992) Collective challenges and collective accomplishments. In: Bostron RP, Watson RT, Kinney ST (eds) Computer augmented team work: A guided tour. Van Nostrand Reinhold, New York, pp 299–324
52. Rouwette EAJA (2003) Group model building as mutual persuasion. Ph D Dissertation, Radboud University Nijmegen
53. Rouwette EAJA, Vennix JAM (2006) System dynamics and organizational interventions. Syst Res Behav Sci 23(4):451–466
54. Rouwette EAJA, Vennix JAM (2007) Team learning on messy problems. In: London M, Sessa VI (eds) Work group learning: Understanding, improving & assessing how groups learn in organizations. Lawrence Erlbaum Associates, Mahwah, pp 243–284
55. Rouwette EAJA, Vennix JAM, van Mullekom T (2002) Group model building effectiveness: A review of assessment studies. Syst Dyn Rev 18(1):5–45
56. Rouwette EAJA, Größler A, Vennix JAM (2004) Exploring influencing factors on rationality: A literature review of dynamic decision-making studies in system dynamics. Syst Res Behav Sci 21(4):351–370
57. Rouwette EAJA, Vennix JAM, Felling AJA (2008) On evaluating the performance of problem structuring methods: An attempt at formulating a conceptual framework. Paper accepted for Group Decision and Negotiation
58. Schein EH (1987) Process consultation, vol II. Addison-Wesley, Reading
59. Shields M (2001) An experimental investigation comparing the effects of case study, management flight simulator and facilitation of these methods on mental model development in a group setting. In: Hines JH, Diker VG, Langer RS, Rowe JI (eds) Proceedings of the 20th international conference of the system dynamics society, Atlanta
60. Shields M (2002) The role of group dynamics in mental model development. In: Davidsen PI, Mollona E, Diker VG, Langer RS, Rowe JI (eds) Proceedings of the 20th international conference of the system dynamics society, Palermo
61. Stenberg L (1980) A modeling procedure for public policy. In: Randers J (ed) Elements of the system dynamics method. Productivity Press, Cambridge, pp 292–312
62. Sterman JD (1994) Learning in and about complex systems. Syst Dyn Rev 10(2–3):291–330
63. van den Belt M (2004) Mediated modeling. A system dynamics approach to environmental consensus building. Island Press, Washington
64. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester
65. Vennix JAM (1999) Group model building: Tackling messy problems. Syst Dyn Rev 15(4):379–401
66. Vennix JAM, Gubbels JW, Post D, Poppen HJ (1990) A structured approach to knowledge elicitation in conceptual model-building. Syst Dyn Rev 6(2):31–45
67. Vennix JAM, Scheper W, Willems R (1993) Group model-building: What does the client think of it? In: Zepeda E, Machuca J (eds) The role of strategic modelling in international competitiveness. Proceedings of the 1993 international system dynamics conference Mexico, Cancun, pp 534–543
68. Vennix JAM, Akkermans HA, Rouwette EAJA (1996) Group model-building to facilitate organizational change: An exploratory study. Syst Dyn Rev 12(1):39–58

69. Vennix JAM, Andersen DF, Richardson GP (eds) (1997) Special issue on group model building. Syst Dyn Rev 13(2):187–201
70. Verburgh LD (1994) Participative policy modelling: Applied to the health care industry. Ph D Dissertation, Radboud University Nijmegen
71. Warren K (2000) Competitive strategy dynamics. Wiley, Chichester
72. Warren K (2005) Improving strategic management with the fundamental principles of system dynamics. Syst Dyn Rev 21(4):329–350
73. Watt CH (1977) Why won't anyone believe us? Simulation 28:1–3
74. Weil HB (1980) The evolution of an approach for achieving implemented results from system dynamic projects. In: Randers J (ed) Elements of the system dynamics method. MIT, Cambridge
75. Winch GW (1993) Consensus building in the planning process: Benefits from a "hard" modelling approach. Syst Dyn Rev 9(3):287–300
76. Wolstenholme EF (1992) The definition and application of a stepwise approach to model conceptualisation and analysis. Eur J Oper Res 59:123–136
77. Zagonel AA (2004) Reflecting on group model building used to support welfare reform in New York state. Unpublished doctoral dissertation, SUNY, Albany

Books and Reviews

Meadows DH, Richardson J, Bruckmann G (1982) Groping in the dark: The first decade of global modelling. Wiley, Chicester

Schwartz RM (1994) The skilled facilitator: Practical wisdom for developing effective groups. Jossey-Bass, San Francisco

Health Care in the United Kingdom and Europe, System Dynamics Applications to

Eric Wolstenholme[1,2]
[1] South Bank University, London, UK
[2] Symmetric SD, Brighton, UK

Article Outline

This paper describes the application of system dynamics to health and social care in Europe.

Systems thinking and the simulation tool set of system dynamics are introduced together with an overview of current strategic health issues and responses in the UK and Europe. A case study is then presented to demonstrate how effective and apposite system dynamics studies can be. This is followed by a pan-European review of applications of system dynamics in epidemiology and in health treatment and diagnosis in different sectors of health and social care, based on an extensive bibliography. Reference is also made to health workforce planning studies. Lastly, a review of future directions is described.

The knowledge base of this paper is located in published work by internal and external consultants and Universities, but it should also be said that there is far more work in system dynamics in health than is referred to in these sources. Many internal and external consultancies undertake studies which remain unpublished.

The description of the subject and the applications described are comprehensive, but the review is a personal interpretation of the current state of a fast-moving field by the author and apologies are made in advance for any unintended omissions.

The case study in Sect. "A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges" is extracted from material published by Springer-Verlag, US and published with their permission.

Glossary

System dynamics

System A collection of elements brought together for a purpose and whose sum is greater than the parts.

Systems thinking The process of interpreting the world as a complex, self regulating and adaptive system.

System dynamics A method based on quantitative computer simulation to enhance learning and policy design in complex systems.

Qualitative system dynamics The application of systems thinking and system dynamics principles, without formal simulation.

Dynamic complexity The number of interacting elements contained in a system and the consequences of their interactions over time.

Human activity system Any system created and regulated by human intervention.

Reductionism The opposite of systemic – seeing the world only in its constituent parts.

Feedback Feedback refers to the interaction of the elements of the system where a system element, X, affects another system element, Y, and Y in turn affects X perhaps through a chain of causes and effects. Feedback thus controls the performance of the system. Feedback can be either natural or behavioral (created by human intervention) (System Dynamics Society).

Unintended consequences Undesirable consequences arising well intended action – or vice versa.

Continuous simulation The aggregate method of computer simulation used in system dynamics based on a continuous time analogy with fluid dynamics and used to test out patterns of behavior over time.

System structure The term used in system dynamics to refer to the total structure of a system (composing processes, organization boundaries, information feedback, policy and delays).

System behavior The term used in system dynamics to refer to the behavior over time of a particular structure.

Reference mode of behavior An observed past trend and future projected trends used to assist defining model scope and time frame.

Discrete entity simulation A method of simulation based on the movement of individual entities through systems over time either as processes or as interactions between entities.

Health and Social Care

Epidemiology The study of factors affecting the health and the incidence and prevalence of illness of populations.

Health treatment The application of drugs, therapies, and medical/surgical interventions to treat illness.

National health service (NHS) The organization in the UK responsible for the delivery of health care.

Primary care trusts (PCTs) The local operating agencies of the NHS, which both commission (buy) and deliver health services.

General practitioners (GPs) Locally-based general clinicians who deliver primary care services and control access to specialist health services.

Social services In England, care services which provide non-health related care, mainly for children and older people, located within local government in the UK.

Nursing/residential home care In England, private and public residential establishments for the care of older people.

Domiciliary care In England, care for older people in their own homes.

Acute hospitals Hospital dealing with short term conditions requiring mainly one-off treatment.

Outliers Patients located in hospital in wards not related to their condition, due to bed capacity issues.

Intermediate care Short term care to expedite the treatment of non-complex conditions.

Definition of the Subject

All too often complexity issues are ignored in decision making simply because they are just too difficult to represent. Managers feel that to expand the boundaries of the decision domain to include intricate, cross-boundary interconnections and feedback will detract from the clarity of the issue at stake. This is particularly true when the interconnections are behavioral and hard to quantify. Hence, the focus of decision making is either very subjective or based on simple, linear, easy to quantify components. However, such a reductionist stance, which ignores information feedback (for example, the effects of health supply on health demand management) and multiple-ownership of issues can result in unsustainable, short term benefits with major unintended consequences.

System dynamics is a particular way of thinking and analyzing situations, which makes visible the dynamic complexity of human activity systems for decision support.

It is particularly important in the health and social care field where there are major issues of complexity associated with the incidence and prevalence of disease, an aging population, a profusion of new technologies and multiple agencies responsible for the prevention and treatment of illness along very long patient pathways. Health is also linked at every stage to all facets of life and health policy has a strong political dimension in most countries.

Introduction

This paper describes and reviews work in applying system dynamics to issues of health and social care in the UK and Europe. Although the fundamental issues in health and social care and many of the strategies adopted are similar the world over, there are differences in culture, operational policies and funding even over short geographical distances. Additionally, the health field can be dissected in many different ways both internally and between countries.

There is, moreover, a fundamental dilemma at the center of health that determines both its structure and emphasis. Although the real long term and systemic solution to better health lies in the prevention of illness, the health field focuses on the study of the incidence and prevalence of disease (Epidemiology) and on the 'health service' issues of how to manage ill health (Health Diagnosis and Treatment).

There are many reasons for this, not the least being that illness prevention is in fact the province of a field much bigger than health, which includes economics, social deprivation, drugs, poverty, power and politics.

The field of system dynamics in health reflects this dilemma. Whilst all studies would conclude that prevention is better than the cure, the majority of applications focus on illness. Whilst more studies are required on the truly systemic goal of moving attention away from the status quo, for example, modeling the German system of health care and drug addicts [52], the major focus and impact of system dynamics in Europe in recent years has been in terms of Epidemiology and Health Treatment. Hence, it is these categories that will be the focus of this paper. However, work often transcends the two and models often include both disease and treatment states. For example, work on AIDS covers both prevalence and drug treatment and work on long term conditions, particularly mental health conditions, covers condition progression as well as alternative therapies.

It is important to emphasize what this paper does not cover. By definition system dynamics is a strategic approach aimed at assisting with the understanding of high level feedback effects at work in organizations. It is therefore separate from the many applications of spreadsheets and discrete entity simulation methods applied to answer short term operational level issues in health [9,21,29].

It is also important to note where the knowledge base of this paper is located. System dynamics applications in health in Europe began in the 1980s and are expanding rapidly. However, as will be seen from the bibliography to this paper, much of the work is applied by internal and external consultants and Universities for health care managers and reported in management, operational research and system dynamics journals. Little of the work so far has been addressed directly at clinicians or published in the health literature. It should also be said that there is far more work in system dynamics in health than is referred to in this publication. Many internal and external consultancies undertake studies which remain unpublished.

Initially the fundamentals of system dynamics will be described followed by an overview of current health issues and responses in the UK and Europe. This is followed by a case study to demonstrate how effective and apposite system dynamics studies can be. There then follows a review of applications in epidemiology and in both physical and mental health diagnosis and treatment. Mention is also made of health workforce planning studies. Lastly, a review of future directions is described.

The History of System Dynamics

System dynamics was conceived at MIT, Boston in the late 60s and has now grown into a major discipline [25,47] which was formally celebrated and reviewed in 2008 [48]. It is widely used in the private business sector in production, marketing, oil, asset management, financial services, pharmaceuticals and consultancy. It is also used in the public sector in defense, health and criminal justice.

System dynamics has a long history in the UK and Europe. The first formal university group was established at the University of Bradford in England 1970. Today there are at least a dozen university departments and business schools offering courses in system dynamics and numerous consultancies of all types using the method in one form or another. Thousands of people have attended private and university courses in system dynamics and, additionally, there are almost one hundred UK members of the System Dynamics Society, which is the largest national grouping outside the US.

The Need for System Dynamics

Most private and public organizations are large and complex. They exhibit both 'detailed' complexity (the number of elements they contain), but more importantly 'dynamic' complexity (the number of interconnections and interactions they embrace). They have long processes which transcend many sectors, each with their own accounting and performance measures. In the case of health and social care organizations this translates into long patient pathways across many agencies. Complexity and decision making in the public sector is also compounded by a multitude of planning time horizons and the political dimension.

Long processes mean that there are many opportunities for intervention, but that the best levers for overall improvement are often well away from symptoms of problems. Such interventions may benefit sectors other than those making the investments and require an open approach to improving patient outcomes, rather than single agency advantage.

The management of complex organizations is complicated by the fact that human beings have limited cognitive ability to understand interconnections and consequently have limited mental models about the structure and dynamics of organizations.

A characteristic of complex organizations is a tendency for management to be risk averse, policy resistant and quick to blame. This usually means they prefer to stick to traditional solutions and reactive, short term gains. In doing this managers ignore the response of other sectors and levels of the organization. In particular, they underestimate the role and effect of behavioral feedback.

Such oversight can result in unintended consequences in the medium term that undermine well-intended actions. Self organizing and adaptive responses in organizations can lead to many types of informal coping actions, which in turn, inhibit the realization of improvement attempts and distort data. A good example of these phenomena, arising from studies described here, is the use of 'length of stay' in health and social care services as a policy lever to compensate for capacity shortages.

Planning within complex organization reflects the above characteristics. The core of current planning tends to be static in nature, sector-based and reliant on data and financial spreadsheets with limited transparency of assumptions. For example the planning of new acute hospitals can quickly progress to detailed levels without assessment of trends in primary and post acute care; that is, where hospital patients come from and go to.

In contrast, sustainable solutions to problems in complex organizations often require novel and balanced interventions over whole processes, which seem to defy logic and may even be counterintuitive.

However, in order to realize such solutions requires a leap beyond both the thinking and planning tools commonly used today. In order to make significant changes in complex organizations it is necessary to think differently and test ideas before use. System dynamics provides such a method.

The Components of System Dynamics

System dynamics is based on the idea of resisting the temptation to be over reactive to events, learning instead to view patterns of behavior in organizations and ground these in the structure (operational processes and policies) of organizations. It uses purpose-built software to map processes and policies at a strategic level, to populate these maps with data and to simulate the evolution of the processes under transparent assumptions, polices and scenarios.

System dynamics is founded upon:

- Non linear dynamics and feedback control developed in mathematics, physics and engineering,
- Human, group and organizational behavior developed in cognitive and social psychology and economics,
- Problem solving and facilitation developed in operational research and statistics.

System dynamics provides a set of *thinking* skills and a set of *modeling* tools which underpin the current trend of 'whole systems thinking' in health and social care.

System Dynamics Thinking Skills for the Management of Complex Organizations

In order to understand and operate in complex organizations it is necessary to develop a wide range of thinking skills [45]. The following are summarized after Richmond [42].

- **Dynamic thinking** – The ability to conceptualize how organizations behave over time and how we would like them to behave.
- **System-as-cause thinking** – The ability to determine plausible explanations for the behavior of the organization over time in terms of past actions.
- **Forest thinking** – The ability to see the "big picture" (transcending organizational boundaries).
- **Operational thinking** – The ability to analyze the contribution made to the overall behavior by the interaction of processes, information feedback, delays and organizational boundaries.
- **Closed-loop thinking** – The ability to analyze feedback loops, including the way that results can feedback to influence causes.
- **Quantitative thinking** – The ability to determine the mathematical relationships needed to model cause and effect.
- **Scientific thinking** – The ability to construct and test hypotheses through modeling.

System Dynamics Modeling Tools for Planning in Complex Organizations

A useful way to appreciate the tool set of system dynamics is by a brief comparison with other computer based management tools for decision support.

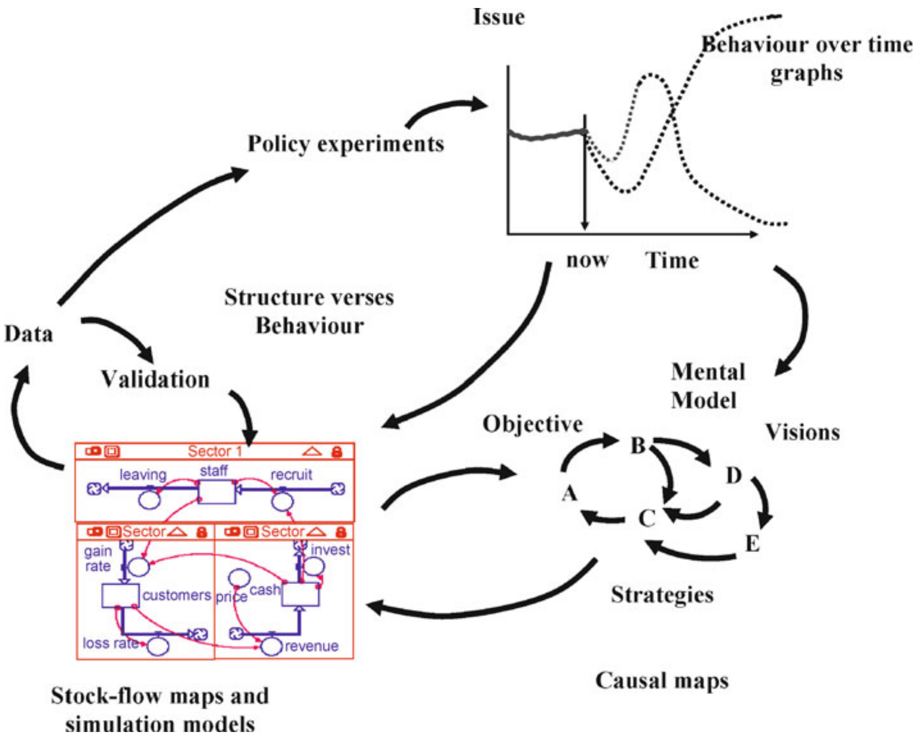
System dynamics is, by definition, a strategic rather than operational tool. It can be used in a detailed operational role, but is first and foremost a *strategic* tool aimed at integrating policies across organizations, where behavioral feedback is important. It is unique in its ability to address the strategic domain and this places it apart from more operational toolsets such as process mapping, spreadsheets, data analysis, discrete entity simulation and agent-based simulation.

System dynamics is based on representing process flows by 'stock' and 'rate' variables. Stocks are important measurable accumulations of physical (and non-physical) resources in the world. They are built and depleted over time as input and output rates to them change under the influence of feedback from the stocks and outside factors. Recognizing the difference between stocks and rates is fundamental to understanding the world as a system. The superimposition of organizational sectors and boundaries on the processes is also fundamental to understanding the impact of culture and power on the flows. System dynamics also makes extensive use of causal maps to both help conceptualize models and to highlight feedback processes within models.

Applying System Dynamics with Management Teams

However, the success of system dynamics lies as much in its process of application as in the tool set and hence demands greater skill in conceptualization and use than spreadsheets.

Figure 1 shows the overall process of applying system dynamics. A key starting point is the definition of an initial significant issue of managerial concern and the estab-



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 1
The systems thinking/system dynamics method

lishment of a set of committed and consistent management teams from all agencies involved in the issue. Another requirement is a set of facilitators experienced in both conceptualizing and formulating system dynamics models. The models created must be shared extensions of the mental models of the management teams, not the facilitators and, importantly owned by the team.

The next step is the analysis of existing trends in major performance measures of the organizations and of their future trajectories, desired and undesired. This is referred to as the reference model of behavior of the issue and helps with the establishment of the time scale of the analysis. The key contribution of system dynamics is then to formulate a high level process map, at an appropriate level of aggregation, linking operations across organizations and to populate this with the best data available. Once validated against past data, the mental models of the management team and shown capable of reproducing the reference mode of behavior of the issue ('what is'), the model is used to design policies to realize desired futures ('what might be'). Maps and models are constructed in relatively inexpensive purpose-built software (for example *ithink*, *Vensim* and *Powersim*) with very transparent graphical interfaces.

The key is to produce the simplest model possible consistent with maintaining its transparency and having confidence in its ability to cast new light on the issue of concern. This means keeping the resolution of the model at the highest possible level and this distinguishes it from most spreadsheets and process maps.

An Overview of Health and Social Care in the UK and Europe

Ensuring that all residents have access to health and social care services is an important goal in all EU countries and all have universal or almost universal health care coverage (European Observatory 'Healthcare in Transition' profiles and OECD Health Data 2004). Even in the Netherlands, where only 65% of the population are covered by a compulsory scheme, with voluntary private insurance available to the remainder, only 1.6% of the population are without health insurance.

At the present time, most care in the EU is publicly financed, with taxation and social insurance provide the main sources of funding. Taxation is collected at either the national level or local level, or both and social insurance contributions are generally made by both employees and employers. The role of private insurance varies between countries and generally private insurance is as a supplement to, rather than as a substitute for, the main care system. The exceptions to this are Germany and the Netherlands. Further, people are increasingly required to pay part of the cost of medical care.

The delivery of health and social care is a mixture of public and private with only 10 countries not having any private delivery sector at all.

This paper is primarily concerned with health and social care supply issues. Although the structure and terminology associated with supply varies across the EU the underlying issues tend to be similar between countries. Hence the major issues will be described for England.

Health in England is primarily managed and delivered by the National Health Service (NHS) and is at the center of a modernization agenda, whereby the government sets out a program of change and targets against which the public may judge improved services.

A major mechanism for reform tends to be via frequent changes to organizational structure. The current structure consist of large primary care trusts (PCTs), which both deliver services such as General Practitioner Services (GPs), but also purchase (commission) more specialist services from other agencies, both public and private. A key driver of structural change is to enhance primary care and to take the pressure off acute hospitals (acute is a word used to differentiate short term hospitals from long stay ones). Initiatives here center on providing new services, such as diagnostic and treatment centers and shorter term 'intermediate' care. Emphasis is on bringing the services to the users, patient choice, payment by results (rather than through block contracts) and service efficiency, the latter being driven by target setting and achievement. The government has made reform of public services a key plank in its legislative program and pressure to achieve a broad range of often conflicting targets is therefore immense. However, despite continual increases in funding new initiatives are slow to take effect and the performance and viability of the service is problematic with money often being used to clear deficits rather than generate new solutions.

Social care in England is delivered both by a public sector located with Local Government Social Services Directorates and a private sector. It consists of numerous services to support children and older people. The latter consisting of care homes, nursing homes and domiciliary (at home) care.

Many patient processes, particularly for older people, transcend health and social care boundaries and hence create a serious conflict of process structure and organizational structure, where the relative power of the different agencies is a major determinant of resource allocation [64]. Consequently, emphasis in this paper will be on joint health and social care work.

A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges

In order to give a flavor of the relevance and impact of applying system dynamics to health and social care issues a concise case study will be presented [65,67,70].

Issue

Delayed hospital discharge was an issue which first came onto the UK legislative agenda in late 2001. The 'reference mode' of behavior over time for this situation was that of increasing numbers of patients occupying hospital beds, although they had been declared "medically fit". In March 2002, 4,258 people were "stuck" in hospital and some were staying a long time, pushing up the number of bed days and constituting significant lost capacity.

The government's approach to this issue was to find out who was supposed to "get the patients out" of acute hospitals and threaten them with 'fines' if they did not improve performance. This organization proved to be social services for older people, who are located within the local government sector and who are responsible for a small, but significant, number of older people needing ex-hospital ('post-acute') care packages. Such patients are assessed and packages organized by hospital social workers. There was also pressure on the government from hospitals claiming that some of the problem was due to lack of hospital capacity.

The idea of fines was challenged by the Local Government Association (LGA), which represents the interests of all local government agencies at the national level) who suggested that a 'system' approach should be undertaken to look at the complex interaction of factors affecting delayed hospital discharges. This organization, together with the NHS Confederation (the partner organization representing the interests of the National Health Service organizations at a national level) then commissioned a system dynamics study to support their stance.

The remit was for consultants working with the representatives of the two organizations to create a system dynamics model of the 'whole patient pathway' extending upstream and downstream from the stock of people delayed in hospital, to identify and test other interventions affecting the issue.

Model

A system dynamics model was developed interactively with managers from the LGA and NHS, using national data to simulate pressures in a sample health economy covering primary, acute and post acute care over a 3 year period. The model was driven by variable demand including three winter pressure "peaks" when capacity in each sector was stretched to the limit. Figure 2 shows an overview of the sectors of the model.

The patient flows through the model were broken down into medical flows and surgical with access to the medical and surgical stocks of beds being constrained by bed capacity. The medical flows were mainly emergencies patients and the surgical flows mainly non-emergency 'elective' patients, who came via referral processes and wait lists.

Further, medical patients were broken down into 'fast' and 'slow' streams. The former were the normal patients who had a short stay in hospital and needed few post acute services and the latter the more complex cases (mainly older people), who require a longer stay and hospital and complex onward care packages from social services. This split was because although the slow patients were few in number they constituted most of the people who caused delayed discharges.

The post hospital health and social care services of intermediate care, nursing/residential home care, and domiciliary care were included in the model and were also capacity constrained in terms of the number of care packages they could provide.

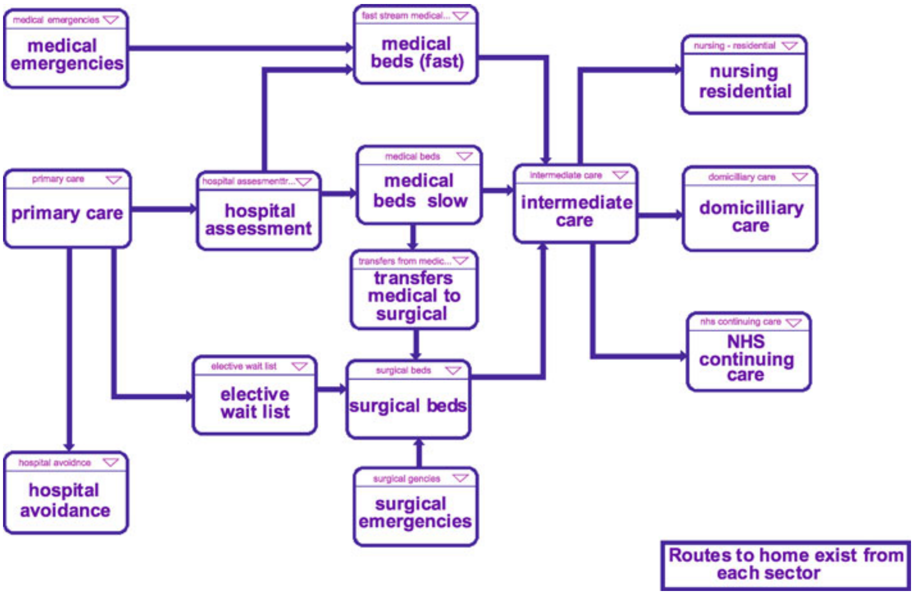
The model incorporated a number of mechanisms by which hospitals coped during periods of high demand, for example, moving medical patients to surgical beds (outliers) and early discharges with allowance for readmissions.

Configuration of the Model

The model was set up to simulate a typical sample health economy over a 3 year period when driven by a variable demand (including three winter "peaks"). The capacity constrained sectors of the model were given barely sufficient capacity to cope. This situation was designed to create shocks against which to test alternative policies for performance improvement. Major performance measures in use in the various agencies were incorporated. These included:

1. Cumulative episodes of elective surgery.
2. Elective wait list size and wait time.
3. Numbers of patients in hospital having completed treatment and assessment, but not yet discharged (delayed discharges).
4. Number of 'outliers'.

The model was initially set up with a number of fixed experiments, to introduce people to the range of experiments



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 2
An overview of the sectors of the delayed discharge model

that yielded useful insights into the behavior of the whole system. From there, they were encouraged to devise their own experiments and develop their own theories of useful interventions and commissioning strategies.

The three main polices tested in the fixed runs were:

1. Adding additional acute hospital bed capacity. This is the classic response used over many years by governments throughout the world to solve any patient pathway capacity problems and was a favorite 'solution' here.
2. Adding additional post acute capacity, both nursing and residential home beds but also more domiciliary capacity.
3. Diverting more people away from hospital admission by use of pre-hospital intermediate capacity and also expansion of treatment in primary care GP surgeries.

Example Results from the Delayed Hospital Discharge Model

Figures 3, 4 and 5 show some typical outputs for the delayed hospital discharge model. Figure 3 captures the way capacity utilization was displayed (actual beds occupied *v* total available for both medical and surgical sectors of the hospital) and shows the occurrence of 'outliers' (transfers of patients from medical to surgical beds) whenever medical capacity was reached.

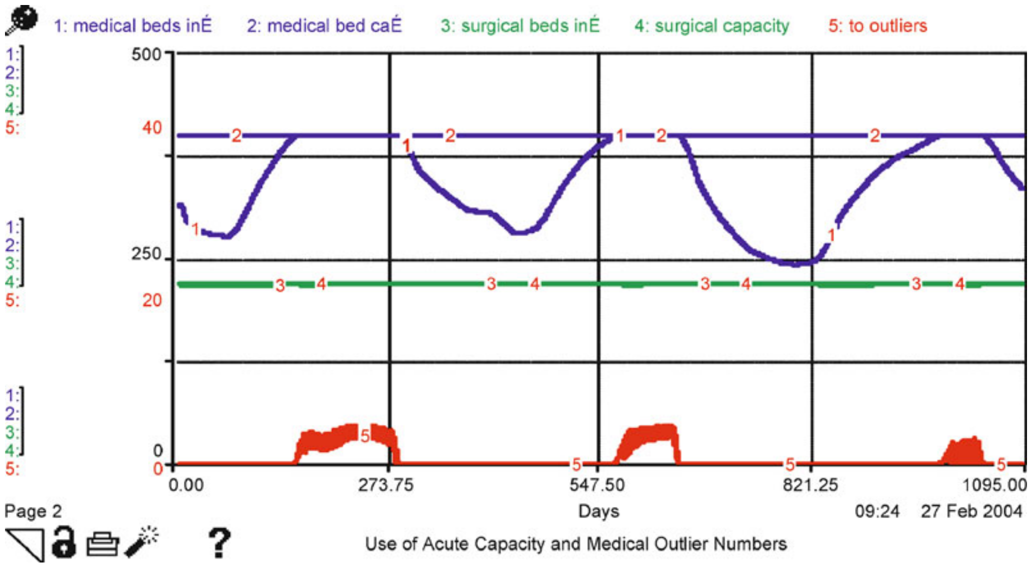
Figures 4 and 5 show comparative graphs of 3 policy runs for 2 major performance measures for 2 sectors of the patient pathway – delayed discharges for post acute social services and cumulative elective procedures for acute hospitals. In each case the base run is line 1. Line 2 shows the effect of increasing hospital beds by 10% and line 3 shows the effect of increasing post acute capacity by 10%.

The interesting feature of this example output is that the cheaper option of increasing post acute capacity gives lower delayed discharges and higher elective operations whereas the more expensive option of increasing acute hospital beds benefits the hospital but makes delayed discharges worse. The key to this counter intuitive effect is that increasing post acute capacity results in higher hospital discharges which in turn reduces the need for the 'outlier' coping policy in the hospital, hence freeing up surgical capacity for elective operations.

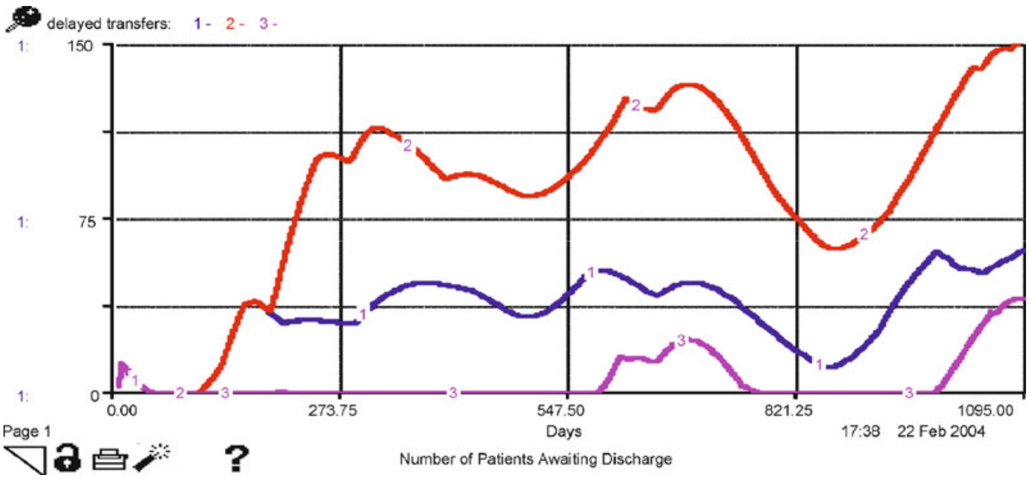
Outcomes

Common Sense Solutions Can Be Misleading The obvious unilateral solution of adding more acute capacity was shown to exacerbate the delayed discharge situation. Increasing hospital capacity means facilitating more hospital admissions, but with no corresponding increase in hospital discharges. Hence, the new capacity will simply fill up and then more early discharges and outliers will be needed.

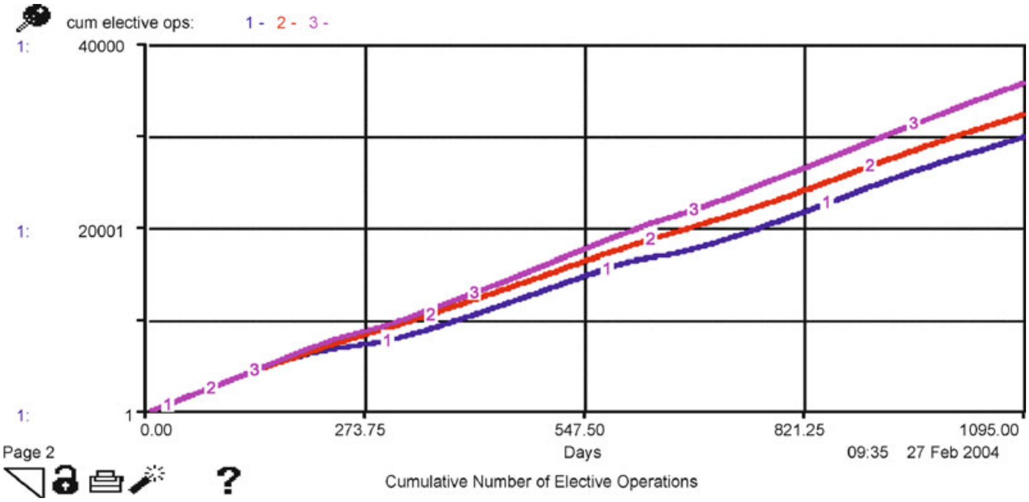
Fines May Have Unintended Consequences This solution was shown to depend on where the money raised by fines was spent. If the money levied from social services



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 3
Medical and surgical bed utilization's in hospital and 'outliers'



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 4
Delayed hospital discharges for 3 policy runs of the model



Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 5
Cumulative elective operations for 3 policy runs of the model

was given to the acute sector to finance additional capacity it was clearly demonstrated that this would make delayed discharges worse. It would be worse still if it causes the post-acute sector to cut services. The effects of service cuts may also then spill over into other areas of local government including housing and education.

It was demonstrated that there were some interventions that could help:

1. Increasing post acute capacity gives a win-win solution to both health and social care because it increases all acute and post acute sector performance measures. Such action allows hospital discharges to directly increase, and eliminates the need for the hospitals to apply coping policies, which in turn increases elective operations and reduces elective wait times. Further, counter intuitively, increasing medical capacity in hospital is more effective than increasing surgical capacity for reducing elective wait times.
2. Reducing assessment times and lengths of stay in all sectors is beneficial to all performance measures, as is reducing variation in flows, particularly reinforcing feedback loops like re-admission rates.
3. Increasing diversion from hospitals into pre-admission intermediate care was almost as beneficial as increasing post acute capacity.
4. If fines are levied they need to be re-invested from a whole systems perspective. This means re-balancing

resources across all the sectors (NOT just adding to hospital capacity).

5. In general the model showed that keeping people out of hospital is more effective than trying to get them out faster. This is compounded by the fact that in-patients are more prone to infections so the longer patients are in hospital, the longer they will be in hospital.
6. Improving the quality of data was shown to be paramount to realizing the benefits of all policies. This is an interesting conclusion associated with many system dynamics studies, where explicit representation of the structure of the organization can lead to a review and redesign of the information needed systems to really manage the organization.

An interesting generalization of the findings was that increasing stock variables where demand is rising (such as adding capacity) is an expensive and unsustainable solution. Whereas increasing rate variables, by reducing delays and lengths of stay, is cheaper and sustainable.

Impact

This model was shown at the Political Conferences of 2002 and generated considerable interest. It was instrumental in causing re-thinking of the intended legislation, so that social services was provided with investment funding to address capacity issues, and the implementation of fines was delayed for a year. Reference to the model was made in the House of Lords.

> Moving the main amendment, Liberal Democrat health spokesperson Lord Clement-Jones asked the House to agree that the Bill failed to tackle the causes of delayed discharges and would create perverse incentives which would undermine joint working between local authorities and the NHS and distort priorities for care of elderly people by placing the requirement to meet discharge targets ahead of measures to avoid hospital admission … **He referred to "ithink", the whole systems approach being put forward by the Local Government Association, health service managers and social services directors involving joint local protocols and local action plans prepared in co-operation.**

Postscript

This case study demonstrates the ability of system dynamics to be applied quickly and purposefully to shed rigor and insight on an important issue. The study enabled the development of a very articulate and compelling case for the government to move from a reactive position of blaming social services to one of understanding and acting on a systemic basis. The whole project including modeling and communication of the outcomes was completed in 6 weeks.

Review of System Dynamics Studies in Epidemiology in Europe

The potential for system dynamics in population health and disease control began in the UK in the late eighties/early nineties with the extensive studies carried out on AIDS modeling. The majority of these studies were by Prof. Brian Dangerfield and Carole Roberts and were ongoing until 2000 [14,15,16,17,18,19,20].

The earlier studies [16] used a transition model to portray the nature of the disease and to better specify the types of data collection required for further developments of the model. The model was then developed further over the years [18,19] and was fed with time-series data of actual cases. This enabled projections of future occurrence to be forecast. The latter models were more concerned with examining the resource and cost implications of treatments given to HIV positive individuals and at their varying stages up until the ensuing onset of AIDS.

A recent study by Dangerfield et al. [20] saw further development of the original model with parameter optimization and recent data on the spread of AIDS in the UK was also integrated. The rationale for the update of the model was to investigate the recent dramatic decrease in diagnosed Aids cases in the West. The model assesses the effects of relatively new emergent triple antiretroviral therapy given to HIV patients causing this reduction and examines the possibility of continuity of the effectiveness of this therapy.

Dangerfield explains some of the reasons [13] why system dynamics acts as an excellent tool for epidemiological modeling. The positive and negative feed-back loops help imitate the natural disposition of the spread and containment of diseases amongst the general population. Further, system dynamics allows delays associated with the incubation predisposition of infectious diseases to be accurately and easily modeled without the need for complicated mathematical representation.

The work in the UK was complemented by work in Holland on simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men [23] and on models for analysis and evaluation of strategies for preventing AIDS [32]. Further epidemiological studies in system dynamics in the UK related to the outbreak out of BSE and the subsequent infection of humans with its human form nvCJD [12].

These models are all characterized by modeling the flow of people through different stocks over time representing the different stages of the disease progression. The purpose of the model is then to test the effects of interventions aimed at slowing down the rate of progression of the condition or indeed moving people 'upstream' to less severe states of the condition.

Review of System Dynamics Studies in Health and Social Care Management in Europe

By far the greatest number of studies and publications in the use of system dynamics in health and social care is associated with patient flow modeling for health care planning. That is, the flow of patients through multiple service delivery channels. Patient pathway definition has been an area of health modernization and these pathways lend themselves to representation as stock/flow resource flows in system dynamics. The purpose of this type of modeling is to identify bottlenecks, plan capacity, reduce wait lists, improve the efficiency of patient assessments and times and the design of alternative pathways with shorter treatment times, (for example, intermediate care facilities both pre and post hospital treatment).

A characteristic of patient flows is that they are long and pass through multiple agencies and hence confront the major health issues of working across boundaries and designing integrated policies. Studies in this area have examined the flow of many different populations of patients and often resulted in arrayed models to represent the flow of different 'populations' or 'needs groups' through several parallel service channels.

The studies have covered both physical and mental conditions and have sometimes combined both the dynamic progression of people through undiagnosed and untreated disease states and the dynamic progression of diagnosed people through treatment pathways.

The Modeling of the Diagnosis and Treatment of Physical Conditions

Here the most common set of models are associated with the flow of patients from primary care, through acute hospitals and onwards into post acute care such as social services provisions for home care, nursing care and residential care. The populations have often been split between the simple everyday cases and the complex cases associated with older people needing greater degrees of care. They have also involves medical and surgical splits. There are a number of review papers which supplement the work described below [1,18,19].

In addition to work in the 1990s on the interface between health and social care [59,60,61] and the national level UK work on older people flows through hospitals [65,67,71,72], Wolstenholme has reported that system dynamics applications are currently underway by the authors in 10 local health communities around the UK with the objectives of modeling patient flows across agency boundaries to provide a visual and quantitative stimulus to strategic multi-agency planning [65].

Lane has reported work in Accident and Emergency Departments [33] and in mapping acute patient flows [34] whilst Royston worked with the NHS to help develop and implement policies and programs in health care in England [43]. Taylor has undertaken award winning modeling of the feedback effects of reconfiguring health services [49,50,51], whilst Lacey has reported numerous UK studies to support the strategic and performance management roles of health service management, including provision of intermediate care and reduction of delayed hospital discharges [31]. Other intermediate care and social care delivery studies are described by Bayer [6,7] and further hospital capacity studies by Coyle [11]. Elsewhere, there have been specific studies on bed-blocking [24] and screening [37].

In Norway system dynamics-based studies have focused on mapping the flows of patients in elderly non-acute care settings [10]. The purpose of this study according to Chen is to differentiate between acute and non-acute settings and thereby increase understanding of the complexity and dynamics caused by influencing elements in the system. Also it is to provide a tool for local communities in Norway for their long term budget planning in the non-acute health sector for the elderly.

Work on reducing waiting lists has been reported in Holland [30,53,54,57]). Also in Holland Vennix has reported comprehensive work on modeling a regional Dutch health care system [56].

Work has been undertaken to balance capacities in individual hospitals in Italy [44] and in Norway [38,41]. Whist normally the realm of more operational types of simulation system dynamics has proved very effective here. There has also been work to assess the impact on health and social care of technological innovation, particularly telecare [5,8]. Additionally, system thinking has been undertaken by doctors to examine the European time directive [40].

Given the similar nature of a lot of these studies further detail here will focus on the work of Vennix in participative model building and Wolstenholme in extracting insights from numerous studies.

Participative Model Building

A characteristic of all Vennix's work has been group model building [55]. The main objectives of this [27] are communication and learning and integration of multiple perspectives where the process of model building is frequently more important than the resulting model itself [56]. Vennix brought together strategic managers and important stakeholders to participate in the process of building a system dynamics model of the Dutch healthcare system. The policy problem which is modeled in Vennix's 1992 study is related to the gradual, but persistent, rise in health care costs in the Netherlands. Vennix [56] attempts to find the underlying causes of those increases that emanate from within the health care system itself rather than focusing on exogenous factors. By doing so Vennix stands to identify potential levers within the health care system that can be practically and appropriately be adjusted to reduce cost increases.

Vennix attempts to extract important assumptions from the key players by posing three straight forward questions;

a) What factors have been responsible for the increase in health care costs?
b) How will health care costs develop in the future?
c) What are the potential effects of several policy options to reduce these costs?

Participants are asked if they agreed or disagreed with the statements and why they thought the statements were true or not. The most frequently given reasons for the verbal statements were then incorporated in to the statements to create causal arguments from the participant's mental models.

Similar methods were adopted to identify policies which represent the aggregate of many individual actions. For example, why a GP may decide on such matters as frequency of patients appointments, drugs choice, referral to other medical specialist or a combination of all these. Vennix's model was subsequently formalized and quantified and converted into a computer-based learning environment for use by a wider range of health personnel.

The idea of using system dynamics as a means of participative modeling for learning is also inherent in other work [35].

Offering Insights into Managing the Demand for Health Care

Wolstenholme reports the insights from many applications of his own and other work. He suggests a hypothesis that the 'normal' mode of operation for many health and social care organizations today is often well beyond their safe design capacity. This situation arises from having to cope with whatever demand arrives at their door irrespective of their supply capability. Risk levels can be high in these organizations and the consequences could be catastrophic for patients [71,72].

Evidence for the hypothesis has emerged at many points along patient pathways in health and social care from a number of studies carried out using system dynamics simulation to identify and promote systemic practice in local health communities. The rigor involved in knowledge-capture and quantitative simulation model construction and running has identified mismatches between how managers claim their organizations work and the observed data and behavior. The discrepancies can only be explained by surfacing informal coping policies. For example, transferring medical emergency patients to surgical wards, resulting in canceled elective procedures, also reported by Lane [35]. Indeed, the data itself becomes questionable as it reflects more the actions of managers than the true characteristics of patients.

The result of capacity pressure can mean that managers are unable, physically and financially, to break out from a fire-fighting mode to implement better resource investment and development policies for systemic and sustainable improvement. The insights reported are important for Health and Social Care management, the meaning of data and for modeling. The key message here is that much-needed systemic solutions and whole system thinking can never be successfully implemented until organizations are allowed to articulate and dismantle their worst coping strategies and return to working within best practice capacities.

The Modeling of the Treatment of Mental Health Diagnosis and Treatments in the UK

Modeling to assist mental health reform has recently developed as a separate strand of health work in the UK [46, 69,72].

Mental health services in the UK over the past 50 years have undergone numerous major reforms. The National Institute for Clinical Excellence [36] has recently published extensive research-based guidelines on the way stepped care might be best achieved. These involved moves towards a balanced, mixed community/institutional provision of services set within a range of significant reforms to the National Health Service. The latest and perhaps most significant reform is that associated with the introduction of 'stepped care'. Stepped care is aimed at bringing help to more patients more cheaply by developing intermediate staff, services and treatments between GPs and the specialist health hospitals.

Having decided on the new treatments at each step and having designed the basic patient pathways, modeling has been used in the North West of England to help with communication of the benefits and to overcome anticipated problems with resource reallocation issues [69]. Further work in Lincolnshire UK [58] reports the increasing use of 'matrix' modeling in mental health to capture the dynamics of both patient needs and treatments. This work also demonstrates the dangers of over-investment in situations where much demand is in accrued backlogs and incidence is reducing due to better and more successful interventions.

The depression work has also led to work at the Department of Health in the UK to help analyze the national impact of stepped services for mental health on the totality of the labor market and unemployment [72]. This work is an example of the value that system dynamics can add to conventional cost benefit analysis. A static cost benefit analysis was developed into a system dynamics model. By developing a bigger picture of the issue, both upstream to where patients go after treatment and downstream from where patients originate in the labor market, and by simulation of the enhanced vision, the dynamic cost benefit analysis is shown to advance understanding of the issue and plans.

The work questions the magnitude of the potential benefits, introduces phasing issues, surfaces structural insights, takes account of the dynamics of the lab-our market and forces linkages between the plan and other initiatives to get people back to work. The paper suggests that cost benefit analysis and system dynamics are very complementary and should be used together in strategic planning.

Other mental health capacity planning studies have been carried out for individual mental health hospitals and trusts. One such study [71] describes the application of system dynamics to assist decision making in the reallocation of resources within a specialist mental health trust in south London. Mental health service providers in the UK are under increasing pressure to both reduce their own costs and to move resources upstream in mental health patient pathways to facilitate treating more people, whilst not compromising service quality.

The investigation here focused on the consequences of converting an existing specialist service ward in a mental health hospital into a 'triage' ward, where patients are assessed and prioritized during a short stay for either discharge or onward admission to a normal ward. Various policies for the transition were studied together with the implications for those patients needing post hospital services and relocation within the community. The model suggested that the introduction of a triage ward could meet the strategic requirement of a 10% shift away from institutional care and into community services. The paper includes a number of statements from the management team involved on the benefits of system dynamics and the impact of its application on their thinking.

System Dynamics Workforce Planning Models to Support Health Management

It is also important to mention that work has been carried out in a number of countries in the field of workforce planning related to health. In the UK the NHS has deployed sophisticated workforce planning models to determine the training and staffing needs associated with numerous alternative service configurations. In the Spanish Health system modeling has been used to determine the number of doctors required for a number of specialists services and to attempt to explore solutions for the current imbalance among supply and demand of physicians [2,4,5]. Elsewhere the factors affecting staff retention has been studied [28] and in the Netherlands, an advisory body of the Dutch government was given the responsibility of implying a new standard for the number of rheumatologists [39]. One of the main factors that were studied in the scenario analysis stage was the influences of changing demographics on the demand of manpower in the health system. Other studies have covered time reduction legislation on doctor training [22].

Future Directions

System dynamics has already made a significant impact on health and social care thinking across the EU. Many policy insights have been generated and the organizations are increasingly being recognized as complex adaptive systems. However, true understanding and implementation of the messages requires much more work and too many organizations are still locked into a pattern of short-termism which leads them to focus on the things they feel able to control – usually variables within their own individual spheres of control. There are also some aspects of system reform in some countries that are producing perverse incentives which encourage organizations to apply short-term policies.

Wider communication of existing studies and further studies are necessary to demonstrate the advantages of sustainable, systemic solutions. The key challenge lies in demonstrating to a wider audience of managers and clinicians that they can add value to the whole whilst remaining autonomous. An important element is to train more people capable of modeling and facilitating studies and to simplify the process and software of system dynamics.

Acknowledgments

The author would like to acknowledge the many practitioners of system dynamics in health and social care throughout Europe for their dedication to using these methods to improve health and wellbeing. Particular thanks are due to my colleagues in the International Systems Dynamics community and its Society and those in the UK Chapter of the Society, without whom I would not have the knowledge base to undertake this review. Special thanks are also due to my colleagues in Symmetric SD without whom much of the personal work reported here would not have been carried out at all.

Bibliography

Primary Literature

1. Abdul-Salam O (2006) An overview of system dynamics applications. In: A dissertation submitted to the University of Salford Centre for Operational Research and Applied Statistics for the degree of MSc Centre for Operational Research and Applied Statistics, University of Salford
2. Alonso Magdaleno MI (2002) Administrative policies and MIR vacancies: Impact on the Spanish Health System. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
3. Alonso Magdaleno MI (2002) Dynamic analysis of some proposals for the management of the number of physicians in Spain. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
4. Alonso Magdaleno MI (2002) Elaboration of a model for the management of the number of specialized doctors in the spanish health system. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
5. Bayer S (2001) Planning the implementation of telecare services. In: The 19th International Conference of the System Dynamics Society, System Dynamics Society, Atlanta, 2001
6. Bayer S (2002) Post-hospital intermediate care: Examining assumptions and systemic consequences of a health policy prescription. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
7. Bayer S, Barlow J (2003) Simulating health and social care delivery. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
8. Bayer S, Barlow J (2004) Assessing the impact of a care innovation: Telecare. In: 22nd International Conference of the System Dynamics Society, System Dynamics Society, Oxford, 2004
9. Brailsford SC, Lattimer VA (2004) Emergency and on-demand health care: Modelling a large complex system. J Operat Res Soc 55:34–42
10. Chen Y (2003) A system dynamics-based study on elderly non-acute service in norway. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
11. Coyle RG (1996) A systems approach to the management of a hospital for short-term patients. Socio Econ Plan Sci 18(4):219–226
12. Curram S, Coyle JM (2003) Are you vMad to go for surgery? Risk assessment for transmission of vCJD via surgical instruments: The contribution of system dynamics. In: Proceedings of the 21st International Conference of the System Dynamics Society, New York, 2003
13. Dangerfield BC (1999) System dynamics applications to european health care issues. System dynamics for policy, strategy and management education. J Operat Res Soc 50(4):345–353
14. Dangerfield BC, Roberts CA (1989) A role for system dynamics in modelling the spread of AIDS. Trans Inst Meas Control 11(4):187–195
15. Dangerfield BC, Roberts CA (1989) Understanding the epidemiology of HIV infection and AIDS: Experiences with a system dynamics. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Simulation Councils Inc, San Diego, pp 241–247
16. Dangerfield BC, Roberts CA (1990) Modelling the epidemiological consequences of HIV infection and AIDS: A contribution from operational research. J Operat Res Soc 41(4):273–289
17. Dangerfield BC, Roberts CA (1992) Estimating the parameters of an AIDS spread model using optimisation software: Results for two countries compared. In: Vennix JAM, Faber J, Scheper WJ, Takkenberg CA (eds) System Dynamics. System Dynamics Society, Cambridge, pp 605–617
18. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five european countries. In: Dangerfield BC, Roberts CA (eds) O.R. Work in HIV/AIDS 2nd edn. Operational Research Society, Birmingham, pp 7–13
19. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: Some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
20. Dangerfield BC, Roberts CA, Fang Y (2001) Model-based scenarios for the epidemiology of HIV/AIDS: The consequences of highly active antiretroviral therapy. Syst Dyn Rev 17(2):119–150
21. Davies R (1985) An assessment of models in a health system. J Operat Res Soc 36:679–687
22. Derrick S, Winch GW, Badger B, Chandler J, Lovett J, Nokes T (2005) Evaluating the impacts of time-reduction legislation on junior doctor training and service. In: Proceedings of the 23rd International Conference of the System Dynamics Society, Boston, 2005
23. Dijkgraaf MGW, van Greenstein GJP, Gourds JLA (1998) Interactive simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men in the Netherlands: A proposal. In: Jager JC, Rotenberg EJ (eds) Statistical Analysis and Mathematical Modelling of AIDS. OUP, Oxford, pp 112–122
24. El-Darzi E, Vasilakis C (1998) A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in hospitals. Health Care Manag Sci 1(2):143–149
25. Forrester JW (1961) Industrial Dynamics. MIT Press

26. Gonzalez B, Garcia R (1999) Waiting lists in spanish public hospitals: A system dynamics approach. Syst Dyn Rev 15(3):201–224
27. Heyne G, Geurts JL (1994) DIAGNOST: A microworld in the healthcare for elderly people. In: 1994 International System Dynamics Conference. System Dynamics Society, Sterling
28. Holmstroem P, Elf M (2004) Staff retention and job satisfaction at a hospital clinic: A case study. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
29. Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. J Operat Res Soc 50(2):109–123
30. Kim DH, Gogi J (2003) System dynamics modeling for long term care policy. Proceedings of the 21st International Conference of the System Dynamics Society. System Dynamics Society, New York
31. Lacey P (2005) Futures through the eyes of a health system simulator, Paper presented to the System Dynamics Conference, Boston 2005
32. Lagergren M (1992) A family of models for analysis and evaluation of strategies for preventing AIDS. In: Jager JC (eds) Scenario Analysis. Elsvier, Amsterdam, pp 117–145
33. Lane DC (2000) Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. J Operat Res Soc 51(5):518
34. Lane DC, Husemann E (2008) System dynamics mapping of acute patient flows. J Oper Res Soc 59:213–224
35. Lane DC, Monefeldt C et al (2003) Client involvement in simulation model building: Hints and insights from a case study in a London hospital. Health Care Manag Sci 6(2):105–116
36. National Institute for Clinical Excellence (2004) Depression: Management of depression in primary and secondary care – NICE guidance. National Clinical Practice Guideline 23
37. Osipenko L (2006) System dynamics model of a new prenatal screening technology (Screening pregnant women). In: 24th International Conference of the System Dynamics Society, Nijmegen, The Netherlands, 23–27 July 2006
38. Petersen LO (2000) How should the capacity for treating heart decease be expanded? In: 18th International Conference of the System Dynamics Society, Bergen, 2000
39. Posmta TJBM, Smits MT (1992) Personnel planning in health care: An example in the field of rheumatology. In: Proceedings of the 1992 International System Dynamics Conference of the System Dynamics Society. System Dynamics Society, Utrecht
40. Ratnarajah M (2005) European union working time directive. In: 7th Annual Gathering. System Dyanamics Society, Harrogate, February 2005
41. Ravn H, Petersen LO (2007) Balancing the surgical capacity in a hospital. Int J Healthcare Technol Manag 14:4023–4089
42. Richmond B (1994) ISEE Systems Inc, Hanover
43. Royston G, Dost A (1999) Using system dynamics to help develop and implement policies and programmes in health care in England. Syst Dyn Rev 15(3):293–315
44. Sedehi H (2001) HDS: Health department simulator. In: The 19th International Conference of the System Dynamics Society. System Dynamics Society, Atlanta, 2001
45. Senge P (1990) The fifth discipline doubleday. New York
46. Smith G, Wolstenholme EF (2004) Using system dynamics in modeling health issues in the UK. In: 22nd International Conference of the System Dynamics Society. The System Dynamics Society, Oxford, 2004
47. Sterman J (2000) Business dynamics: System thinking and modelling for a complex world. McGraw-Hill, Boston
48. Sterman T (ed) (2008) Exploring the next frontier: System Dynamics at 50. Syst Dyn Rev 23:89–93
49. Taylor KS (2002) A system dynamics model for planning and evaluating shifts in health services: The case of cardiac catheterisation procedures in the NHS London, London School of Economics and Political Science. J Opr Res Soc 56:659–1229
50. Taylor KS, Dangerfield BC (2004) Modelling the feedback effects of reconfiguring health services. J Operat Res Soc 56:659–675 (Published on-line Sept 2004)
51. Taylor KS, Dangerfield BC, LeGrand J (2005) Simulation analysis of the consequences of shifting the balance of health care: A system dynamics approach. J Health Services Res Policy 10(4):196–202
52. Tretter F (2002) Modeling the system of health care and drug addicts. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
53. Van Ackere A, PC Smith (1999) Towards a macro model of national health service waiting lists. Syst Dyn Rev 15(3):225
54. Van Dijkum C, Kuijk E (1998) Experiments with a non-linear model of health-related actions. In: 16th International Conference of the System Dynamics Society. System Dynamics Society, Quebec, 1998
55. Vennix AM (1996) Group Model Building. Wiley, Chichester
56. Vennix JAM, JW Gubbels (1992) Knowledge elicitation in conceptual model building: A case study in modeling a regional dutch health care system. Europ J Operat Res 59(1):85–101
57. Verburgh LD, Gubbels JW (1990) Model-based analyses of the dutch health care system. In: System Dynamics '90: Proceedings of the 1990 International Systems Dynamics Conference, International System Dynamics Society, Chestnut Hill, 1990
58. Wolstenholme E, McKelvie D, Monk D, Todd D, Brad C (2008) Emerging opportunities for system dynamics in UK health and social care – the market-pull for systemic thinking. Paper submitted to the 2008 System Dynamics Conference, Athens 2008
59. Wolstenholme EF (1993) A case study in community care using systems thinking. J Operat Res Soc 44(9):925–934
60. Wolstenholme EF (1996) A management flight simulator for community care. In: Cropper S (ed) Enhancing decision making in the NHS. Open University Press, Milton Keynes
61. Wolstenholme EF (1999) A patient flow perspective of UK health services: Exploring the case for new intermediate care initiatives. Syst Dyn Rev 15(3):253–273
62. Wolstenholme EF (2004) Using generic system archetypes to support thinking and learning. Syst Dyn Rev 20(2):341–356
63. Wolstenholme EF, McKelvie D (2004) Using system dynamics in modeling health and social care commissioning in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
64. Wolstenholme EF, Monk D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
65. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling health and social care commissioning in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004

66. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling mental health issues in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
67. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
68. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2005) Coping but not coping in health and social care – masking the reality of running organisations well beyond safe design capacity. Syst Dyn Rev 23(4):371–389
69. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2006) Reforming mental health services in the UK – using system dynamics to support the design and implementation of a stepped care approach to depression in north west england. In: Proceedings of the 2006 System Dynamics Conference, Nijmegen, 2006
70. Wolstenholme EF, Monk D, McKelvie D (2007) Influencing and interpreting health and social care policy in the UK In: Qudrat-Ullah H, Spector MJ, Davidsen PI (eds) Complex decision making: Theory and practice. Springer, Berlin
71. Wolstenholme EF, Monk D, McKelvie D, Gillespie P, O'Rourke D, Todd D (2007) Reallocating mental health resources in the borough of Lambeth, London, UK. In: Proceedings of the 2007 System Dynamics Conference, Boston, 2007
72. Wolstenholme EF, Monk D, McKelvie D, Todd D (2007) The contribution of system dynamics to cost benefit analysis – a case study in planning new mental health services in the UK. In: Proceedings of the 2007 System Dynamics Conference, Boston

Books and Reviews

Dangerfield BC (1999) System dynamics applications to european health care issues. J Operat Res Soc 50(4):345–353

Dangerfield BC, Roberts CA (eds) (1994) Health and health care dynamics. Special issue of Syst Dyn Rev 15(3)

Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

Health Care in the United States, System Dynamics Applications to

Gary Hirsch[1], Jack Homer[2]
[1] Independent Consultant, Wayland, USA
[2] Independent Consultant, Voorhees, USA

Article Outline

Glossary

Chronic illness A disease or adverse health state that persists over time and cannot in general be cured, although its symptoms may be treatable.

Stock An accumulation or state variable, such as the size of a population.

Flow A rate-of-change variable affecting a stock, such as births flowing into a population or deaths flowing out.

Feedback loop A closed loop of causality that acts to counterbalance or reinforce prior change in a system state.

Definition of the Subject

Health care involves a complex system of interactions among patients, providers, payers, and other stakeholders. This system is difficult to manage in the United States because of its free market approach and relative lack of regulation. System Dynamics simulation modeling is an effective method for understanding and explaining causes of dysfunction in U.S. health care and for suggesting approaches to improving health outcomes and slowing rising costs. Applications since the 1970s have covered diverse areas in health care including the epidemiology of diseases and substance abuse, as well as the dynamics of health care capacity and delivery and their impacts on health. Many of these applications have dealt with the mounting burden of chronic illnesses, such as diabetes. In this article four such applications are described.

Introduction

Despite remarkable successes in some areas, the health enterprise in the United States faces difficult challenges in meeting its primary goal of reducing the burden of disease and injury. These challenges include the growth of the underinsured population, epidemics of obesity and asthma, the rise of drug-resistant infectious diseases, ineffective management of chronic illness [33], long-standing racial and ethnic health disparities [32], and an overall decline in the health-related quality of life [64]. Many of these complex problems have persisted for decades, often proving resistant to attempts to solve them [36].

It has been argued that these interventions fail because they are made in piecemeal fashion, rather than comprehensively and from a whole-system perspective [15]. This compartmentalized approach is engrained in the financial structures, intervention designs, and evaluation methods of most health agencies. Conventional analytic methods are generally unable to satisfactorily address situations in which population needs change over time (often in response to the interventions themselves), and in which risk factors, diseases, and health resources are in a continuous state of interaction and flux [52].

The term *dynamic complexity* has been used to describe such evolving situations [56]. Dynamically complex problems are often characterized by long delays between causes and effects, and by multiple goals and interests that may in some ways conflict with one another. In such situations, it is difficult to know how, where, and when to intervene, because most interventions will have unintended consequences and will tend to be resisted or undermined by opposing interests or as a result of limited resources or capacities.

The systems modeling methodology of System Dynamics (SD) is well suited to addressing the challenges of dynamic complexity in public health. The methodology involves the development of causal diagrams and policy-oriented computer simulation models that are unique to each problem setting. The approach was developed by computer pioneer Jay W. Forrester in the mid-1950s and first described at length in his book *Industrial Dynamics* [11] with some additional principles presented in later works [8,9,10,12]. The International System Dynamics Society was established in 1983, and within the Society a special interest group on health issues was organized in 2003.

SD modeling has been applied to health and health care issues in the U.S. since the 1970s. Topic areas have included:

- Disease epidemiology including work in heart disease [24,40], diabetes [24,34,43], obesity [25], HIV/AIDS [29], polio [57] and drug-resistant pneumococcal infections [28];
- Substance abuse epidemiology covering heroin addiction [37], cocaine prevalence [30], and tobacco reduction policy [50,58];

- Health care capacity and delivery in such areas as population-based HMO planning [21], dental care [20,38], and mental health [38], and as affected by natural disasters or terrorist acts [16,22,41]; and
- Interactions between health care or public health capacity and disease epidemiology [17,18,19,23,27].

Most of these modeling efforts have been done with the close involvement of clinicians and policymakers who have a direct stake in the problem being modeled. Established SD techniques for group model building [60] can help to harness the insights and involvement of those who deal with public health problems on a day-to-day basis.

It is useful to consider how SD models compare with those of other simulation methods that have been applied to public health issues, particularly in epidemiological modeling. One may characterize any population health model in terms of its degree of aggregation, that is, the extent to which individuals in the population are combined together in categories of disease, risk, or age and other demographic attributes. At the most aggregate end of the scale are lumped contagion models [3,35]; more disaggregated are Markov models [13,31,44]; and the most disaggregated are microsimulations at the level of individuals [14,51,63].

The great majority of SD population health models are high or moderately high in aggregation. This is related to the fact that most SD models have a broad model boundary sufficient to include a variety of realistic causal factors, policy levers, and feedback loops. Although it is possible to build models that are both broad in scope and highly disaggregated, experience suggests that such very large models nearly always suffer in terms of their ability to be easily and fully tested, understood, and maintained. In choosing between broader scope and finer disaggregation, SD modelers tend to opt for the former, because a broad scope is generally needed for diagnosing and finding effective solutions to dynamically complex problems [55,56].

The remainder of this article describes four of the System Dynamics modeling applications cited above, with a focus on issues related to chronic illnesses and their care and prevention. The U.S. Centers for Disease Control and Prevention (CDC) estimates that chronic illness is responsible for 70% of all deaths and 75% of all health care costs in the U.S. [5]. The applications discussed below address:

- Diabetes and heart failure management at the community level;
- Diabetes prevention and management from an epidemiological perspective;
- General chronic illness care and prevention at a community level; and
- General chronic illness care and prevention at the national level.

The article concludes with a discussion of promising areas for future work.

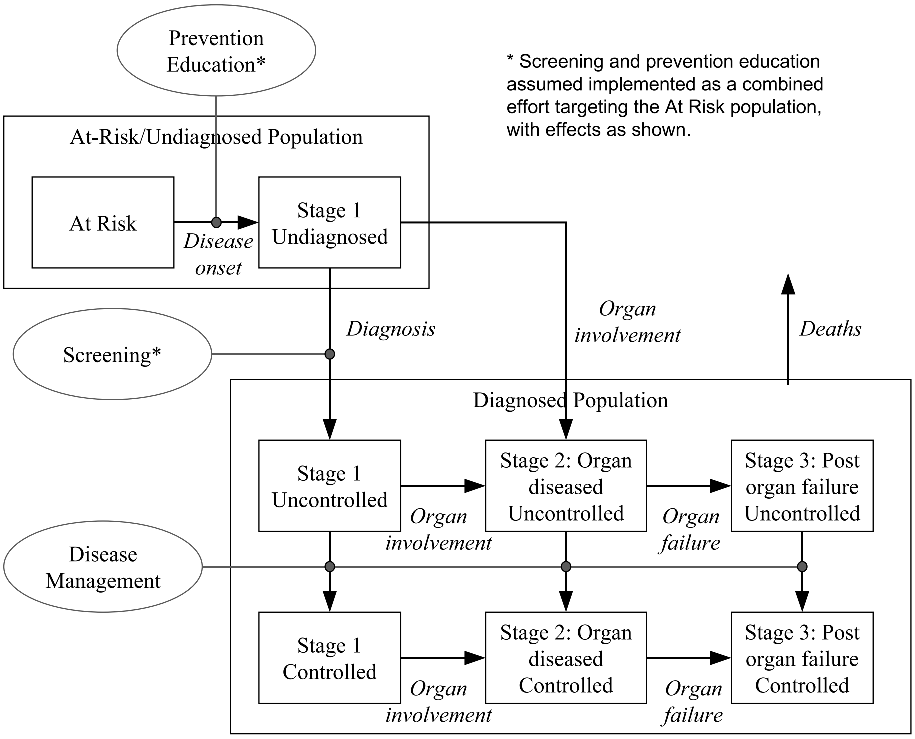
Four Applications

Diabetes and Heart Failure Management at the Community Level

Two hours north of Seattle in the state of Washington lies Whatcom County, with a population of about 170 thousand. The county embarked on a major effort to address chronic illness care and was selected by the Robert Wood Johnson Foundation as one of seven sites in a larger chronic care program called Pursuing Perfection [24]. The program initially concentrated on two chronic illnesses as prototypes for improved care: diabetes and congestive heart failure. Both of these illnesses affect millions of people in the U.S. and other countries and exact a heavy toll in terms of direct medical expenditures as well as indirect costs due to disability and premature mortality [2,45,47]. The prevalence of both diseases is growing rapidly as the numbers of people above age 65 increase, and also due to the epidemic rise in obesity, which is a risk factor for both diabetes and heart disease [7,46].

Leaders of the Whatcom County program had two critical needs for making decisions about potential interventions for improving the care of chronic illnesses such as diabetes and heart failure. First, they wanted to get a sense of the overall impact of these interventions on incidence and prevalence of diabetes and heart failure, health care utilization and cost, and mortality and disability rates in the community. Second, they wanted to understand the impact of the various interventions on individual health care providers in the community and on those who pay for care—insurers, employers, and patients themselves. There was a concern that the costs and benefits of the program be shared equitably and that providers who helped produce savings should not suffer a resulting loss of revenue to their businesses.

These analytic needs could not be met with spreadsheet and other models that project impacts in a simple, linear fashion. Interventions in chronic illness do not have simple direct impacts. The aging of the population, incidence of new cases, progression of disease, deaths, and the interventions themselves all create a constantly changing situation. Interventions ideally reduce mortality rates, but



* Screening and prevention education assumed implemented as a combined effort targeting the At Risk population, with effects as shown.

Health Care in the United States, System Dynamics Applications to, Figure 1
Disease stages and intervention points in the Whatcom County Diabetes Model

this leaves more people with the disease alive and requiring care for years to come.

Figure 1 presents a simplified view of the stock-and-flow structure used in modeling non-insulin-dependent (Type 2) diabetes. The actual model has two separate structures like those shown in Fig. 1, one for the 18-to-64 age group and one for the 65-and-older age group, which are linked by flows of patients turning 65. The model also calculates an inflow of population turning 18, death outflows from each stock based on patient age and stage of illness, and flows of migration into and out of the county. The rectangular boxes in Fig. 1 represent sub-populations with particular characteristics. The arrows signify flows of people from one population group to another (e. g., from uncontrolled to controlled diabetes at a particular stage). Lines from ovals (programmatic interventions such as disease management) to population flows indicate control of or influence on those flows.

The three stages of diabetes portrayed in this figure were identified through discussions with clinicians in Whatcom County. The population At Risk includes those with family history, the obese, and, most directly, those with a condition of moderate blood sugar known as prediabetes. Further increases in blood sugar lead to Stage 1 diabetes, in which blood vessels suffer degradation, but there is not yet any damage to organs of the body, nor typically any symptoms of the encroaching disease. More than half of Stage 1 diabetics are undiagnosed. If Stage 1 diabetics go untreated, most will eventually progress to Stage 2, marked by organ disease. In Stage 2 diabetes, blood flow disturbances impair the functioning of organ systems and potentially lead to irreversible damage. A patient who has suffered irreversible organ damage, or organ failure, is said to be in Stage 3; this would include diabetics who suffer heart attacks, strokes, blindness, amputations, or endstage renal disease. These patients are at the greatest risk of further complications leading to death.

Several studies have demonstrated that the incidence, progression, complications, and costs of diabetes can be reduced significantly through concerted intervention [1,4,6,59,61]. Such intervention may include primary prevention or disease management. As indicated in Fig. 1, primary prevention would consist of efforts to screen the at-risk population and educate them about the diet and activity changes they need to prevent progression to diabetes. Disease management, on the other hand, addresses existing diabetics. A comprehensive disease management approach, such as that employed by the Whatcom County program, can increase the fraction of patients who are able to keep their blood sugar under effective control from the 40% or less typically seen without a program up to perhaps 80% or more.

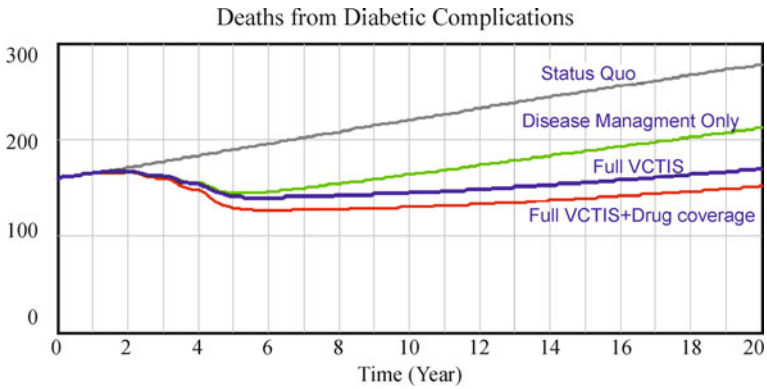
The SD model of diabetes in Whatcom County was first used to produce a 20-year status quo or baseline projection, which assumes that no intervention program is implemented. In this projection, the prevalence of diabetes among all adults gradually increases from 6.5% to 7.5%, because of a growing elderly population; the prevalence of diabetes among the elderly is 17%, compared with 5% among the non-elderly. Total costs of diabetes, including direct costs for health care and pharmaceuticals and indirect economic losses due to disability, grow substantially in this baseline projection.

The next step was to use the model to examine the impact of various program options. These included: (1) a partial approach enhancing disease management but not primary prevention, (2) a full implementation approach combining enhancement of both disease management and primary prevention, and (3) an approach that goes beyond full implementation by also providing greater financial assistance to the elderly for purchasing drugs needed for the control of diabetes.

Simulations of these options projected results in terms of various outcome variables, including deaths from complications of diabetes and total costs of diabetes. Figure 2 shows typical simulation results obtained by projecting these options, in this case, the numbers of deaths over time that might be expected due to complications of diabetes. “Full VCTIS” refers to the complete program of primary prevention and disease management. Under the status quo projection, the number of diabetes-related deaths grows continuously along with the size of the diabetic population. The partial (disease management only) approach is effective at reducing deaths early on, but becomes increasingly less effective over time. The full program approach (including primary prevention) overcomes this shortcoming and by the end of the 20 year simulation reduces diabetes-related deaths by 40% relative to the status quo. Addition of a drug purchase plan for the elderly does even better, facilitating greater disease control and thereby reducing diabetes related deaths by 54% relative to the status quo.

With regard to total costs of diabetes, the simulations indicate that the full program approach can achieve net savings only two years after the program is launched. Four years after program launch, a drug plan for the elderly generates further reductions in disability costs beyond those provided by the program absent such a plan. The partial program approach, in contrast, achieves rapid net savings initially, but gives back most of these savings over time as diabetes prevalence grows. By the end of 20 years, the full program approach results in a net savings amounting to 7% of the status quo costs, two-thirds of that savings coming from reduction in disability-related costs. The model suggests that these anticipated net savings are the result of keeping people in the less severe stages of the diseases for a longer period of time and reducing the number of diabetes-related hospitalizations.

The simulations provided important information and ideas to the Whatcom County program planners, as well as supporting detailed discussions of how various costs and benefits could be equitably distributed among the participants. This helped to reassure participants that none of them would be unfairly affected by the proposed chronic illness program. Perhaps the most important contribution of modeling to the program planning process was its ability to demonstrate that the program, if implemented in its full form, would likely reduce total costs, even though it



Health Care in the United States, System Dynamics Applications to, Figure 2
Typical results from policy simulations with Whatcom County Diabetes Model

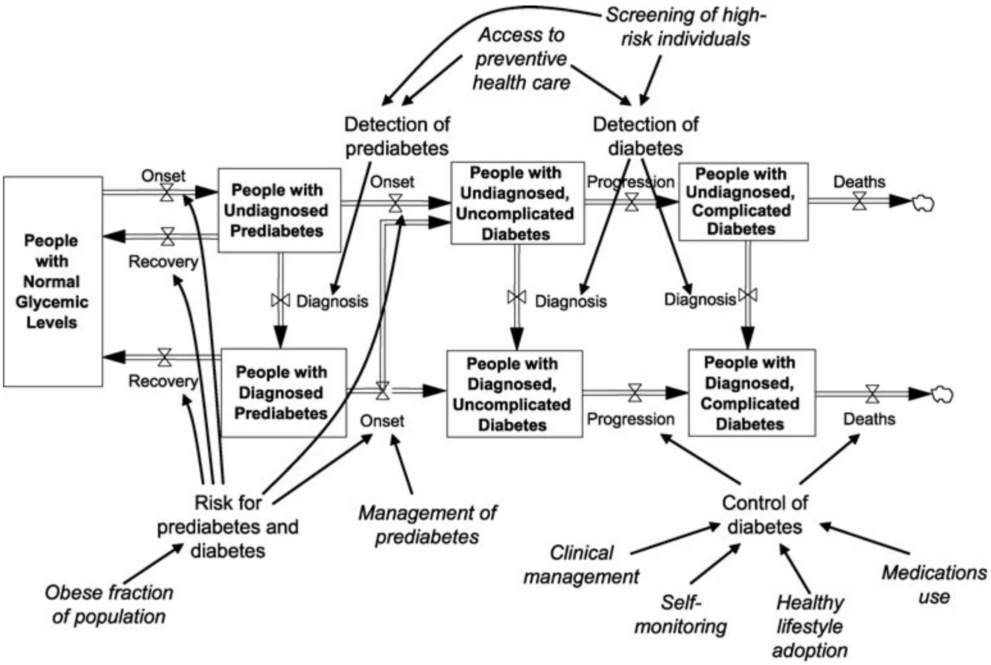
would extend the longevity of many diabetics requiring costly care. Given the sensitivity of payers who were already bearing high costs, this finding helped to motivate their continued participation in the program.

Diabetes Prevention and Management from an Epidemiological Perspective

Another SD model of diabetes in the population was developed for the CDC's Division of Diabetes Translation [34]. This model, a structural overview of which is presented in Fig. 3, builds upon the Whatcom County work but looks more closely at the drivers of diabetes onset, including the roles of prediabetes and obesity. The core of the CDC model is a chain of population stocks and flows portraying the movement of people among the stages of normal blood glucose, prediabetes, uncomplicated diabetes, and complicated diabetes. The prediabetes and diabetes stages are further divided among stocks of people whose conditions are diagnosed or undiagnosed. Also shown in Fig. 3 are the potentially modifiable influences in the model that affect the rates of population flow. These flow-rate drivers include obesity and the detection and management of prediabetes and of diabetes.

The model's parameters were calibrated based on historical data available for the U.S. adult population, as well as estimates from the scientific literature. The model is able to reproduce historical time series, some going as far back as 1980, on diagnosed diabetes prevalence, the diagnosed fraction of diabetes, prediabetes prevalence, the obese fractions of people with prediabetes and diabetes, and the health burden (specifically, the mortality, morbidity, and costs) attributable to diabetes. The model suggests that two forces worked in opposition to affect the diabetes health burden from 1980 to 2004. The first force is a rise in the prevalence of obesity, which led to a greater incidence and prevalence of prediabetes and diabetes through the chain of causation seen in Fig. 3. The second and opposing force is a significant improvement in the control of diabetes, achieved through greater efforts to detect and manage the disease. The second force managed to hold the health burden of diabetes more or less flat during 1980 to 2004.

Looking toward the future, a baseline scenario assumes that no further changes occur in obesity prevalence after 2006, and that inputs affecting the detection and management of prediabetes and diabetes remain fixed at their 2004 values through 2050. This fixed-inputs assumption for the baseline scenario is not meant to represent a forecast of what is most likely to in the future but does provide a useful and easily-understood starting point for policy analysis.


Screening of high-risk individuals
Access to preventive health care
Detection of prediabetes
Detection of diabetes
People with Normal Glycemic Levels
Onset
Recovery
People with Undiagnosed Prediabetes
Onset
People with Undiagnosed, Uncomplicated Diabetes
Progression
People with Undiagnosed, Complicated Diabetes
Deaths
Diagnosis
Diagnosis
Diagnosis
Recovery
People with Diagnosed Prediabetes
Onset
People with Diagnosed, Uncomplicated Diabetes
Progression
People with Diagnosed, Complicated Diabetes
Deaths
Risk for prediabetes and diabetes
Management of prediabetes
Control of diabetes
Obese fraction of population
Clinical management
Self-monitoring
Healthy lifestyle adoption
Medications use

Health Care in the United States, System Dynamics Applications to, Figure 3
Structure of the CDC Diabetes Model

The baseline simulation indicates a future for diabetes burden outcomes for the period 2004–2050 quite different from the past. With obesity prevalence fixed, by assumption, at a high point of 37% from 2006 onward, the diabetes onset rate remains at a high point as well, and diabetes prevalence consequently continues to grow through 2050, becoming more level (after about 2025) only when the outflow of deaths starts to catch up with the inflow of onset.

The CDC model has been used to examine a variety of future scenarios involving policy interventions (singly or in combination) intended to limit growth in the burden of diabetes. These include scenarios improving the management of diabetes, increasing the management of prediabetes, or reducing the prevalence of general population obesity over time. Enhanced diabetes management can significantly reduce the burden of diabetes in the short term, but does not prevent the growth of greater burden in the longer term due to the growth of diabetes prevalence. Indeed, the effect of enhanced diabetes management on diabetes prevalence is not to decrease it at all, but rather to increase it somewhat by increasing the longevity of people with diabetes. Increased management of prediabetes does, in contrast, reduce diabetes onset and the growth of diabetes prevalence. However, it does not have as much impact as one might expect; this is because many people with prediabetes are not diagnosed, and also because the policy does nothing to reduce the growth of prediabetes prevalence due to obesity in the general population. A reduction in prediabetes can be achieved only by reducing population obesity. Significant obesity reduction may take 20 years or more to accomplish fully, but the model suggests that such a policy can be quite a powerful one in halting the growth of diabetes prevalence and burden even before those 20 years are through.

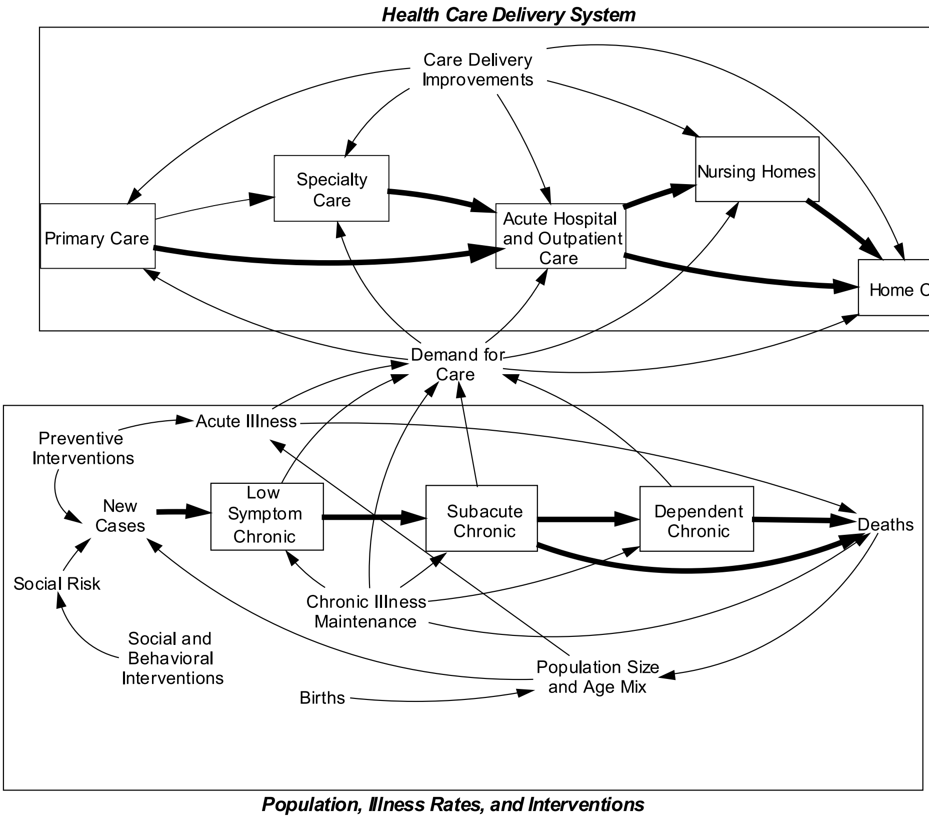
Overall, the CDC model suggests that no single type of intervention is sufficient to limit the growth of the diabetes burden in both the short term and the long term. Rather, what is needed is a combination of disease management for the short term and primary prevention for the longer term. The model also suggests that effective primary prevention may require obesity reduction in the general population a focus on managing diagnosed prediabetes.

At the state and regional level, the CDC model has become the basis for a model-based workshop called the "Diabetes Action Lab". Participants have included state and local public health officials along with non-governmental stakeholders including health care professionals, leaders of not-for-profit agencies, and advocates for people living with diabetes. The workshops have helped the participants improve their intervention strategies and goals and become more hopeful and determined about seeing their actions yield positive results in the future.

The CDC diabetes model has led to other SD modeling efforts at the CDC emphasizing disease prevention, including studies of obesity [25] and cardiovascular risk. The obesity study involved the careful analysis of population survey data to identify patterns of weight gain over the entire course of life from childhood to old age. It explored likely impacts decades into the future of interventions to reduce or prevent obesity that may be targeted at specific age categories. Tentative findings included (1) that obesity in the U.S. should be expected to grow at a much slower pace in the future than it did in the 1980s and 1990s; (2) that the average amount of caloric reduction necessary to reverse the growth of obesity in the population is less than 100 calories per day; (3) that the current trend of focusing intervention efforts on school-age children will likely have only a small impact on future obesity in the adult population; and (4) that it may take decades to see the full impacts of interventions to reduce obesity in the overall population.

General Health Care and Illness Prevention at a Community Level

Hirsch and Immediato [19] describe a comprehensive view of health at the level of a community. Their "Health Care Microworld", depicted in highly simplified form in Fig. 4, simulates the health status and health care delivery for people in the community. The Microworld was created for a consortium of health care providers who were facing a wide range of changes in the mid-1990s and needed a means for their staffs to understand the implications of those changes for how they managed. The underlying SD model consists of many hundreds of equations and was designed to reflect with realistic detail a typical American community and its providers, with data taken from public sources as well as proprietary surveys. Users of the Microworld have a wide array of options for expanding the capacity and performance of the community's health care delivery system such as adding personnel and facilities, investing in clinical information systems, and process redesign. They have a similar range of alternatives for improving health status and changing the demand for care including screening for and enhanced maintenance care of people with chronic illnesses, programs to reduce behavioral risks such as smoking and alcohol abuse, environmental protection, and longer-term risk reduction strategies such as providing social services, remedial education, and job training.

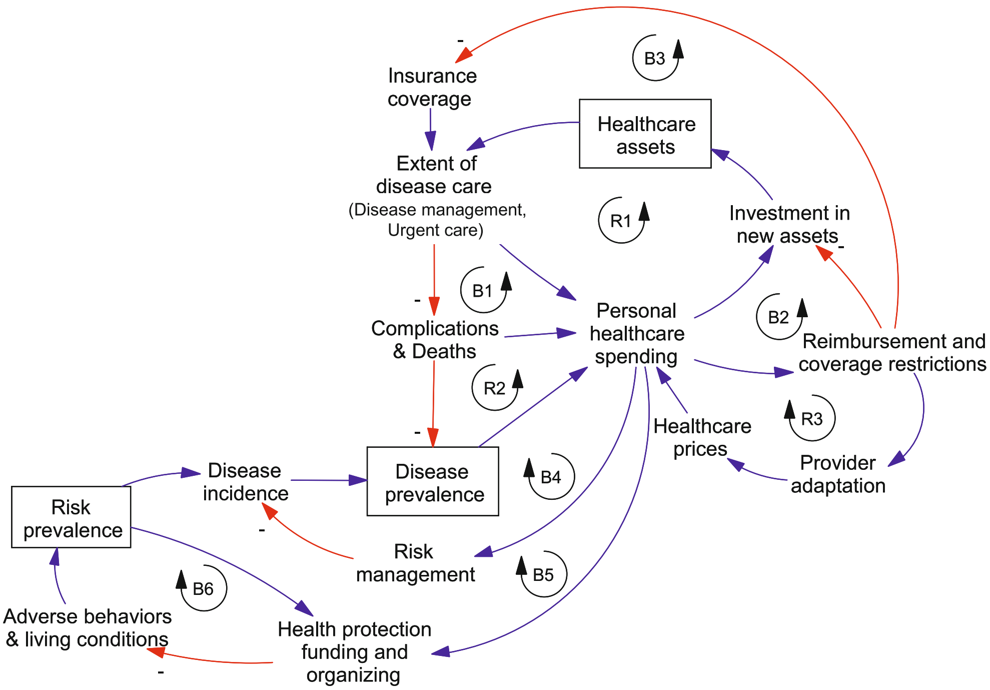

Health Care Delivery System
Care Delivery Improvements
Specialty Care
Nursing Homes
Primary Care
Acute Hospital and Outpatient Care
Home Care
Demand for Care
Acute Illness
Preventive Interventions
New Cases
Low Symptom Chronic
Subacute Chronic
Dependent Chronic
Deaths
Social Risk
Chronic Illness Maintenance
Social and Behavioral Interventions
Population Size and Age Mix
Births
Population, Illness Rates, and Interventions

Health Care in the United States, System Dynamics Applications to, Figure 4
Overview of the Health Care Microworld

The Microworld's comprehensive view of health status and health care delivery can provide insights not available from approaches that focus on one component of the system at a time. For example, users can play roles of different providers in the community and get a better understanding of why many attempts at creating integrated delivery systems have failed because participating providers care more about their own bottom lines and prerogatives than about creating a viable system. When examining strategies for improving health status, users can get a better sense of how a focus on enhanced care of people with chronic illnesses provides short-term benefits in terms of reduced deaths, hospital admissions, and costs, but how better long-term results can be obtained by also investing in programs that reduce social and behavioral health risks.

General Health Care and Illness Prevention at the National Level

Despite rapid growth in health care spending in the U.S. in recent decades, the health of Americans has not noticeably improved. A recent SD model [23] addresses the question of why the U.S. has not been more successful in preventing and controlling chronic illness. This model can faithfully reproduce patterns of change in disease prevalence and mortality in the U.S., but its structure is a generic one and should be applicable to other countries. The model examines the growing prevalence of disease and responses to it, responses which include the treatment of complications as well as disease management activities designed to slow the progression of illness and reduce the occurrence of future complications. The model shows how progress in complications treatment and disease management has slowed since 1980 in the U.S., largely due to a behavioral tug-of-war between health care payers and providers that has resulted in price inflation and an unstable climate for health care investments. The model is also used to demonstrate the impact of moving "upstream" by managing known risk factors to prevent illness onset, and moving even further upstream by addressing adverse behaviors and living conditions linked to the development of these risk factors in the first place.



Health Care in the United States, System Dynamics Applications to, Figure 5

Overview of a National-Level Model of Health Care and Illness Prevention. Key to feedback loops ("R" denotes self-reinforcing, "B" denotes counterbalancing):
R1 Health care revenues are reinvested for further growth
B1 Disease management reduces need for urgent care
R2 Disease care prolongs life and further increases need for care
B2 Reimbursement restriction limits spending growth
B3 Insurance denial limits spending growth
R3 Providers circumvent reimbursement restrictions, leading to a tug-of-war with payers
B4 Risk management proportional to downstream spending can help limit it
B5 Health protection proportional to downstream spending can help limit it
B6 Health protection (via sin taxes) proportional to risk prevalence can help limit it

An overview of the model's causal structure is presented in Fig. 5. The population stock of disease prevalence is increased by disease incidence and decreased by deaths. The death rate can be reduced by a greater extent of disease care, including urgent care and disease management. Disease incidence draws from a stock of risk prevalence, where risk refers to physical or psychological conditions or individual behaviors that may lead to disease. Effective risk management can reduce the flow of people from risk to disease, and may also in some cases allow people to return to a condition of being no longer at risk. Such management may include changes in nutrition or physical activity, stress management, or the use of medications. The risk prevalence stock is increased by adverse behaviors and living conditions. Adverse behaviors may include poor diet, lack of physical activity, or substance abuse. Adverse living conditions can encompass many factors, including crime, lack of access to healthy foods, inadequate regulation of smoking, weak social networks, substandard housing, poverty, or poor educational opportunities.

The extent of care is explained in the model by two key factors: the abundance of health care assets, and insurance coverage. Health care assets are the structures and fixed equipment used directly for health care or for the production of health care products, as well as the human capital of personnel involved. Insurance coverage refers to the fraction of the population with some form of health care insurance, either with a private insurer or through a government plan. The uninsured are less likely than the insured to receive health care services, especially disease management services, something which most of the uninsured cannot afford whereas in most cases they can get urgent care at a hospital emergency department.

The stock of assets is increased by investments, which may be viewed as the reinvestment of some fraction of health care revenues. Such reinvestment drives further growth of care and revenue, and the resulting exponential growth process is identified as loop R1 in Fig. 5. The data indicate, however, that the reinvestment process has slowed significantly since 1980. It is hypothesized that this decline in the reinvestment rate has been the response by potential investors to various forms of cost control, including the restriction of insurance reimbursements, which affect the providers of health care goods and services. With increasing controls and restrictions, these potential investors face greater risk and uncertainty about the future return on their investments, and the result is a greater reluctance to build a new hospital wing, or to purchase an expensive new piece of equipment, or even, at an individual level, to devote a decade or more of one's life to the hardship of medical education and training. Health care costs and cost controls have also led to elimination of private health insurance coverage by some employers, although some of the lost coverage has been replaced by publicly-funded insurance.

One additional part of the downstream health care story portrayed in Fig. 5 is the growth of health care prices. Health care prices are measured in terms of a medical care consumer price index (CPI), which since 1980 has grown much more rapidly than the general CPI for the overall economy. For the period 1980–2004, inflation in medical care prices averaged 6.1% versus general inflation of 3.5%. Why has health care inflation exceeded that of the general economy? Several different phenomena have contributed to health care inflation, but not all have contributed with sufficient magnitude or with the timing necessary to explain the historical pattern. One phenomenon that does appear to have such explanatory power is shown in Fig. 5 as "provider adaptation". This is the idea that, in response to cost containment efforts, providers may "increase fees, prescribe more services, prescribe more complex services (or simply bill for them), order more follow-up visits, or do a combination of these..." [49] Many tests and procedures are performed that contribute little or no diagnostic or therapeutic value, thereby inflating the cost per quality of care delivered. By one estimate, unnecessary and inflationary expense may have represented 29% of all personal health care spending in the year 1989 [23].

The dynamics involving the extent of disease care are portrayed in Fig. 5 in the feedback loops labeled R1, B1, R2, B2, B3, and R3. Taken together, one may view these loops—with the exception of Loop R3—as the story of a "rational" downstream health care system that favors growth and investment until the resulting costs get to a point where further increases are perceived to be no longer worth the expected incremental improvements in health and productivity. Loop R3, however, introduces dysfunction into this otherwise rational system. The loop describes a tug-of-war between payers restricting reimbursement in response to high health care costs, and providers adapting to these restrictions by effectively raising health care prices in an attempt to circumvent the restrictions and maintain their incomes. Because this loop persistently drives up health care costs, it ends up hurting health care investments and insurance coverage (through Loops B2 and B3, respectively), thus dampening growth in the extent of care.

Simulations of the model suggest that there are no easy downstream fixes to the problem of an underperforming and expensive health care system in the U.S. mold. The simulations seem to suggest—perhaps counterintuitively—that health insurance should be stable and non-restrictive in its reimbursements, so as to avoid behavioral backlashes that can trigger health care inflation and under-investment. Although a broad mandate of this sort would likely be politically infeasible in the U.S., movement in this direction could perhaps start with the government's own Medicare and Medicaid insurance programs, and then diffuse naturally to private insurers over time. It is interesting to consider whether a more generous and stable approach to reimbursement could not only combat illness better than the current restrictive approach, but do it more efficiently and perhaps even at lower cost.

The model also includes structure for evaluating the upstream prevention of disease incidence. There are two broad categories of such efforts described in the literature: Risk management for people already at risk, and health protection for the population at large to change adverse behaviors and mitigate unhealthy living conditions. While spending on population-based health protection and risk management programs has grown somewhat, it still represents a small fraction of total U.S. health care spending, on the order of 5% in 2004 [23].

Figure 5 includes three balancing loops to indicate how, in general terms, efforts in risk management and health protection might be funded or resourced more systematically and in proportion to indicators of capability or relative need. Loop B4 suggests that funding for programs promoting risk management could be made proportional to spending on downstream care, so that when downstream care grows funding for risk management would grow as well. Loop B5 suggests something similar for health protection, supposing that government budgets and philanthropic investments for health protection could be set in proportion to recent health care spending. Loop B6 takes a different approach to the funding of health protection, linking it not to health care spending but to risk prevalence, the stock which health protection most directly seeks to reduce. The linkage to risk prevalence can be made fiscally through "sin taxes" on unhealthy items, such as cigarettes (already taxed throughout the U.S. to varying extents [39]) and fatty foods [42]. In theory, the optimal magnitude of such taxes may be rather large in some cases, as the taxes can be used both to discourage unhealthy activities and promote healthier ones [48].

Simulations of the model suggest that whether the approach to upstream action is risk management or health protection, such actions can reduce illness prevalence and ultimately save money. However, the payback time, in terms of reduced downstream health care costs, may be a relatively long one, perhaps on the order of 20 years. It should be noted, however, that the model does not include losses in productivity to employers and society at large. The Whatcom County models described above suggest that when these losses are taken into account, the payback on upstream action may shrink to a much shorter time period that may be acceptable to the public as well as to those decision makers in a position to put upstream efforts into effect [24].

Future Directions

As long as there are dynamically complex health issues in search of answers, the SD approach will have a place in the analytic armamentarium. There is still much to be learned about the population dynamics of individual chronic conditions like hypertension and risk factors like obesity. SD models could also address multiple interacting diseases and risks, giving a more realistic picture of their overall epidemiology and policy implications, particularly where the diseases and risks are mutually reinforcing. For example, it has been found that substance abuse, violence, and AIDS often cluster in the same urban subpopulations, and that such "syndemics" are resistant to narrow policy interventions [53,54,62]. This idea could also be extended to the case of mental depression, which is often exacerbated by other chronic illnesses, and may, in turn, interfere with the proper management of those illnesses. An exploratory simulation model has indicated that SD can usefully address the concept of syndemics [26].

There is also more to be learned about health care delivery systems and capacities, with the inclusion of characteristics specific to selected real-world cases. Models combining delivery systems and risk and disease epidemiology could help policymakers and health care providers understand the nature of coordination required to put ambitious public health and risk reduction programs in place without overwhelming delivery capacities. Such models could reach beyond the health care delivery system per se to examine the potential roles of other delivery systems, such as schools and social service agencies, in health risk reduction.

The more complete view of population health dynamics advocated here may also be extended to address persistent challenges in the U.S. that will likely require policy changes at a national and state level, and not only at the level of local communities. Examples include the large underinsured population, persistent racial and ethnic health disparities, and the persistent shortage of nurses. SD modeling can help to identify the feedback structures responsible for these problems, and point the way to policies that can make a lasting difference.

Bibliography

1. American Diabetes Association and National Institute of Diabetes and Digestive and Kidney Diseases (2002) The prevention or delay of Type 2 diabetes. Diabetes Care 25:742–749
2. American Heart Association (2000) 2001 Heart and Stroke Statistical Update. AHA, Dallas
3. Anderson R (1994) Populations, infectious disease and immunity: A very nonlinear world. Phil Trans R Soc Lond B346:457–505
4. Bowman BA, Gregg EW, Williams DE, Engelgau MM, Jack Jr L (2003) Translating the science of primary, secondary, and tertiary prevention to inform the public health response to diabetes. J Public Health Mgmt Pract November (Suppl):S8–S14
5. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion (2007) Chronic Disease Overview. Available at http://www.cdc.gov/nccdphp/overview.htm
6. Diabetes Prevention Program Research Group (2002) Reduction in the incidence of Type 2 diabetes with lifestyle intervention or metformin. New Engl J Med 346:393–403
7. Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999-2000. JAMA 288:1723–1727
8. Forrester JW, Senge PM (1980) Tests for building confidence in system dynamics models. In: System Dynamics, TIMS Studies in the Management Sciences. North-Holland, New York, pp 209–228
9. Forrester JW (1971) Counterintuitive behavior of social systems. Technol Rev 73:53–68
10. Forrester JW (1980) Information sources for modeling the national economy. J Amer Stat Assoc 75:555–574
11. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
12. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
13. Gunning-Schepers LJ (1989) The health benefits of prevention: A simulation approach. Health Policy Spec Issue 12:1–255
14. Halloran EM, Longini IM, Nizam A, Yang Y (2002) Containing bioterrorist smallpox. Science 298:1428–1432
15. Heirich M (1999) Rethinking Health Care: Innovation and Change in America. Westview Press, Boulder

16. Hirsch GB (2004) Modeling the consequences of major incidents for health care systems. In: 22nd International System Dynamics Conference. System Dynamics Society, Oxford. Available from: http://www.systemdynamics.org/publications.htm
17. Hirsch G, Homer J (2004) Integrating chronic illness management, improved access to care, and idealized clinical practice design in health care organizations: A systems thinking approach. In: International Conference on Systems Thinking in Management. AFEI/University of Pennsylvania, Philadelphia. Available from: http://www.afei.org/documents/CDRomOrderForm.pdf
18. Hirsch G, Homer J (2004) Modeling the dynamics of health care services for improved chronic illness management. In: 22nd International System Dynamics Conference. System Dynamics Society, Oxford. Available from: http://www.systemdynamics.org/publications.htm
19. Hirsch GB, Immediato CS (1999) Microworlds and generic structures as resources for integrating care and improving health. Syst Dyn Rev 15:315–330
20. Hirsch GB, Killingsworth WR (1975) A new framework for projecting dental manpower requirements. Inquiry 12:126–142
21. Hirsch G, Miller S (1974) Evaluating HMO policies with a computer simulation model. Med Care 12:668–681
22. Hoard M, Homer J, Manley W, Furbee P, Haque A, Helmkamp J (2005) Systems modeling in support of evidence-based disaster planning for rural areas. Int J Hyg Environ Health 208:117–125
23. Homer J, Hirsch G, Milstein B (2007) Chronic illness in a complex health economy: The perils and promises of downstream and upstream reforms. Syst Dyn Rev 23(2–3):313–334
24. Homer J, Hirsch G, Minniti M, Pierson M (2004) Models for collaboration: How system dynamics helped a community organize cost-effective care for chronic illness. Syst Dyn Rev 20:199–222
25. Homer J, Milstein B, Dietz W, Buchner D, Majestic E (2006) Obesity population dynamics: Exploring historical growth and plausible futures in the U.S. In: 24th International System Dynamics Conference. System Dynamics Society, Nijmegen. Available from: http://www.systemdynamics.org/publications.htm
26. Homer J, Milstein B (2002) Communities with multiple afflictions: A system dynamics approach to the study and prevention of syndemics. In: 20th International System Dynamics Conference. System Dynamics Society, Palermo. Available from: http://www.systemdynamics.org/publications.htm
27. Homer J, Milstein B (2004) Optimal decision making in a dynamic model of community health. In: 37th Hawaii International Conference on System Sciences. IEEE, Waikoloa. Available from: http://csdl.computer.org/comp/proceedings/hicss/2004/2056/03/2056toc.htm
28. Homer J, Ritchie-Dunham J, Rabbino H, Puente LM, Jorgensen J, Hendricks K (2000) Toward a dynamic theory of antibiotic resistance. Syst Dyn Rev 16:287–319
29. Homer JB, St. Clair CL (1991) A model of HIV transmission through needle sharing. Interfaces 21:26–49
30. Homer JB (1993) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9:49–78
31. Honeycutt AA, Boyle JP, Broglio KR et al (2003) A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050. Health Care Mgmt Sci 6:155–164
32. Institute of Medicine (Board on Health Sciences Policy) (2003) Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. National Academies Press, Washington, DC
33. Institute of Medicine (Committee on Quality of Health Care in America) (2001) Crossing the Quality Chasm: A New Health System for the 21st Century. National Academies Press, Washington, DC
34. Jones AP, Homer JB, Murphy DL, Essien JDK, Milstein B, Seville DA (2006) Understanding diabetes population dynamics through simulation modeling and experimentation. Am J Public Health 96(3):488–494
35. Kaplan EH, Craft DL, Wein LM (2002) Emergency response to a smallpox attack: The case for mass vaccination. In: Proceedings of the Natl Acad of Sciences 99:10935–10940
36. Lee P, Paxman D (1997) Reinventing public health. Annu Rev Public Health 18:1–35
37. Levin G, Roberts EB, Hirsch GB (1975) The Persistent Poppy. Ballinger, Cambridge
38. Levin G, Roberts EB, Hirsch GB, Kligler DS, Wilder JF, Roberts N (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
39. Lindblom E (2006) State cigarette excise tax rates and rankings. Campaign for Tobacco-Free Kids, Washington, DC. Available from: http://www.tobaccofreekids.org/research/factsheets/pdf/0097.pdf
40. Luginbuhl W, Forsyth B, Hirsch G, Goodman M (1981) Prevention and rehabilitation as a means of cost-containment: The example of myocardial infarction. J Public Health Policy 2:1103–1115
41. Manley W, Homer J et al (2005) A dynamic model to support surge capacity planning in a rural hospital. In: 23rd International System Dynamics Conference, Boston. Available from: http://www.systemdynamics.org/publications.htm
42. Marshall T (2000) Exploring a fiscal food policy: The case of diet and ischaemic heart disease. BMJ 320:301–305
43. Milstein B, Jones A, Homer J, Murphy D, Essien J, Seville D (2007) Charting Plausible Futures for Diabetes Prevalence in the United States: A Role for System Dynamics Simulation Modeling. Preventing Chronic Disease, 4(3), July 2007. Available at: http://www\ignorespaces.cdc.gov/pcd/issues/2007/jul/06_0070.htm
44. Naidoo B, Thorogood M, McPherson K, Gunning-Schepers LJ (1997) Modeling the effects of increased physical activity on coronary heart disease in England and Wales. J Epidemiol Community Health 51:144–150
45. National Institute of Diabetes and Digestive and Kidney Diseases (2004) National Diabetes Statistics. Available from: http://diabetes.niddk.nih.gov/dm/pubs/statistics/index.htm
46. National Institute of Diabetes and Digestive and Kidney Diseases (2004) Statistics Related to Overweight and Obesity. Available from: http://www.niddk.nih.gov/health/nutrit/pubs/statobes.htm#preval
47. O'Connell JB, Bristow MR (1993) Economic impact of heart failure in the United States: Time for a different approach. J Heart Lung Transplant 13(suppl):S107-S112
48. O'Donoghue T, Rabin M (2006) Optimal sin taxes. J Public Econ 90:1825–1849
49. Ratanawijitrasin S (1993) The dynamics of health care finance: A feedback view of system behavior. Ph.D. Dissertation, SUNY Albany
50. Roberts EB, Homer J, Kasabian A, Varrell M (1982) A systems

view of the smoking problem: Perspective and limitations of the role of science in decision-making. Int J Biomed Comput 13:69–86

51. Schlessinger L, Eddy DM (2002) Archimedes: A new model for simulating health care systems—the mathematical formulation. J Biomed Inform 35:37–50
52. Schorr LB (1997) Common Purpose: Strengthening Families and Neighborhoods to Rebuild America. Doubleday/Anchor Books, New York
53. Singer M, Clair S (2003) Syndemics and public health: Reconceptualizing disease in bio-social context. Med Anthropol Q 17:423–441
54. Singer M (1996) A dose of drugs, a touch of violence, a case of AIDS: Conceptualizing the SAVA syndemic. Free Inquiry 24:99–110
55. Sterman JD (1988) A skeptic's guide to computer models. In: Grant L (ed) Foresight and National Decisions. University Press of America, Lanham, pp 133–169
56. Sterman JD (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
57. Tompson KM, Tebbens RJD (2007) Eradication versus control for poliomyelitis: An economic analysis. Lancet 369:1363–1371. doi:10.1016/S0140-6736(07)60532-7
58. Tengs TO, Osgood ND, Chen LL (2001) The cost-effectiveness of intensive national school-based anti-tobacco education: Results from the tobacco policy model. Prev Med 33:558–70
59. U.K. Prospective Diabetes Study Group (1998) Tight blood pressure control and risk of macrovascular and microvascular complications in Type 2 diabetes. The Lancet 352:703–713
60. Vennix JAM (1996) Group Model-building: Facilitating Team Learning Using System Dynamics. Wiley, Chichester
61. Wagner EH, Sandhu N, Newton KM et al (2001) Effect of improved glycemic control on health care costs and utilization. JAMA 285(2):182–189
62. Wallace R (1988) A synergism of plagues. Environ Res 47:1–33
63. Wolfson MC (1994) POHEM: A framework for understanding and modeling the health of human populations. World Health Stat Q 47:157–176
64. Zack MM, Moriarty DG, Stroup DF, Ford ES, Mokdad AH (2004) Worsening trends in adult health-related quality of life and self-rated health–United States, 1993-2001. Public Health Rep 119:493–505

Macroeconomics, Non-linear Time Series in

JAMES MORLEY
Washington University, St. Louis, USA

Article Outline

Glossary

Nonlinear time series in macroeconomics A field of study in economics pertaining to the use of statistical analysis of data in order to make inferences about nonlinearities in the nature of aggregate phenomena in the economy.

Time series A collection of data corresponding to the values of a variable at different points of time.

Linear Refers to a class of models for which the dependence between two random variables can be completely described by a fixed correlation parameter.

Nonlinear Refers to the class of models for which the dependence between two random variables has a more general functional form than a linear equation and/or can change over time.

Structural change A change in the model describing a time series, with no expected reversal of the change.

Level Refers to a definition of the business cycle that links the cycle to alternation between phases of expansion and recession in the level of economic activity.

Deviations Refers to a definition of the business cycle that links the cycle to transitory deviations of economic activity from a trend level.

Fluctuations Refers to a definition of the business cycle that links the cycle to any short-run changes in economic activity.

Deepness A characteristic of a process with a skewed unconditional distribution.

Steepness A characteristic of a process with a skewed unconditional distribution for its first-differences.

Sharpness A characteristic of a process for which the probability of a peak when increasing is different than the probability of a trough when decreasing.

Time reversibility The ability to substitute $-t$ and t in the equations of motion for a process without changing the process.

Markov-switching models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled depends on an unobserved discrete Markov process.

Self-exciting threshold models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values.

Nuisance parameters Parameters that are not of direct interest in a test, but influence the distribution of a test statistic.

Pivotal Refers to the invariance of the distribution of a test statistic with respect to values of parameters in the data generating process under the null hypothesis.

Size Probability of false rejection of a null hypothesis in repeated experiments.

Power Probability of correct rejection of a null hypothesis in repeated experiments.

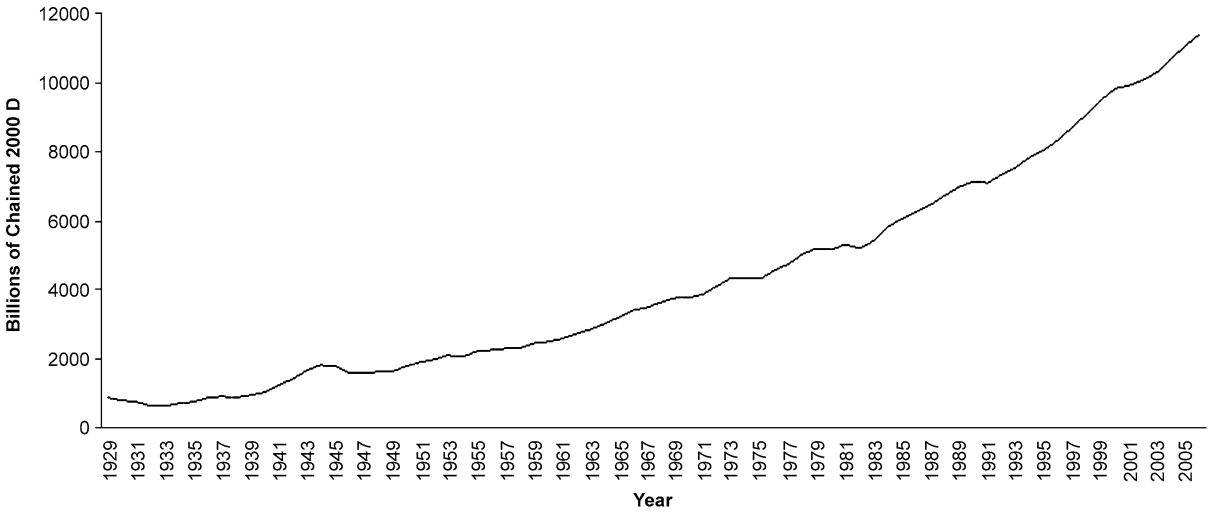
Definition of the Subject

Nonlinear time series in macroeconomics is a broad field of study in economics. It refers to the use of statistical analysis of data to make inferences about nonlinearities in the nature of aggregate phenomena in the economy. This analysis is relevant for forecasting, the formulation of economic policy, and the development and testing of macroeconomic theories.

Introduction

In macroeconomics, the primary aggregate phenomenon is the flow of total production for the entire economy over the course of a year, which is measured by real gross domestic product (GDP). A collection of data corresponding to the values of a variable such as real GDP at different points of time is referred to as a *time series*. Figure 1 presents the time series for US real GDP for each year from 1929 to 2006.

Time series analysis employs stochastic processes to explain and predict the evolution of a time series. In particular, a process captures the idea that different observations are in some way related to each other. The relationship can simply be that the observations behave as if they are drawn from random variables with the same distribution. Or the relationship can be that the distribution assumed to generate one observation depends on the values of other



Macroeconomics, Non-linear Time Series in, Figure 1
US real GDP 1929–2006 (Source: St. Louis Fed website)

observations. Either way, a relationship implies that the observations can be used jointly to make inferences about the parameters describing the distributions (a.k.a. "estimation").

Within the context of time series in macroeconomics, the terms "linear" and "nonlinear" typically refer to classes of models for processes, although other meanings arise in the literature. For the purposes of this survey, a model that assumes the dependence between two random variables in a process can be completely captured by a fixed correlation parameter is said to be *linear*. A very basic example of a linear time series model is the workhorse first-order autoregressive (AR(1)) model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t , \quad \varepsilon_t \sim \text{ i.i.d.}\,(0, \sigma^2)\,, \qquad (1)$$

where $|\phi| < 1$. In words, the random variable y_t that generates the observation in period t is a linear function of the random variable y_{t-1} that generates the observation in period $t - 1$. The process $\{y_t\}_{-\infty}^{\infty}$ is stochastic because it is driven by random "shocks", such as ε_t in period t. These shocks have the same distribution in every period, meaning that, unlike with y_t and y_{t-1}, the distribution of ε_t does not depend on the value of ε_{t-1} or, for that matter, any other shock in any other period (hence the "i.i.d." tag, which stands for "independently and identically distributed"). It is straightforward to show that the correlation between y_t and y_{t-1} is equal to ϕ and this correlation describes the entire dependence between the two random variables. Indeed, for the basic AR(1) model, the dependence and correlation between any two random variables y_t and y_{t-j}, for all t and j, depends only on the fixed parameter ϕ according to the simple function ϕ^j and, given $|\phi| < 1$, the process has finite memory in terms of past shocks. For other time series models, the functions relating parameters to correlations (i. e., "autocorrelation generating functions") are generally more complicated, as are the restrictions on the parameters to ensure finite memory of shocks. However, the models are still linear, as long as the parameters and correlations are fixed.

In contrast to the linear AR(1) model in (1) and other models with fixed correlations, any model that allows for a more general functional form and/or time variation in the dependence between random variables can be said to be *nonlinear*. This nomenclature is obviously extremely open-ended and examples are more revealing than general definitions. Fortunately, macroeconomics provides many examples, with "nonlinear" typically used to describe models that are closely related to linear models, such as the AR(1) model, but which relax one or two key assumptions in order to capture some aspect of the data that cannot be captured by a linear model. The focus of this survey is on these types of nonlinear models.

It should be mentioned at the outset that, in addition to nonlinear models, "nonlinear time series" evokes nonparametric and semiparametric methods (e. g., neural networks). These methods tend to be data intensive and so find more use in finance and other fields where sample sizes are larger than in macroeconomics. "Nonlinear time series" also evokes the development and application of tests for nonlinearity. However, these are the purview of econometrics, not macroeconomics. Thus, tests for nonlinearity will only be discussed in the context of applications that are particularly relevant to the choice of appropriate models for macroeconomic data.

Types of Nonlinear Models

Starting with the linear AR(1) model in (1), there are many ways to introduce nonlinearities. An obvious way is to consider a nonlinear specification for the relationship between the random variables in the model. For example, consider the simple bilinear model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t + \theta(\varepsilon_{t-1} \cdot y_{t-1}) ,\quad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2) . \tag{2}$$

See Granger and Andersen [57] and Rao and Gabr [139] on bilinear models. In macroeconomics at least, there are relatively few applications of bilinear models, although see Peel and Davidson [119], Rothman [128], and Hristova [71].

A more typical approach to introducing nonlinearities in macroeconomics is to allow one (or more) of the parameters in a linear model to be driven by its own process. For example, in a macroeconomics paper that was motivated in part by bilinear models, Engle [46] assumed the squares of shocks (i. e., ε_t^2) follow an AR process, with the implication that the conditional variance of y_t is no longer a constant parameter. Given an AR(1) assumption for ε_t^2, the conditional variance is

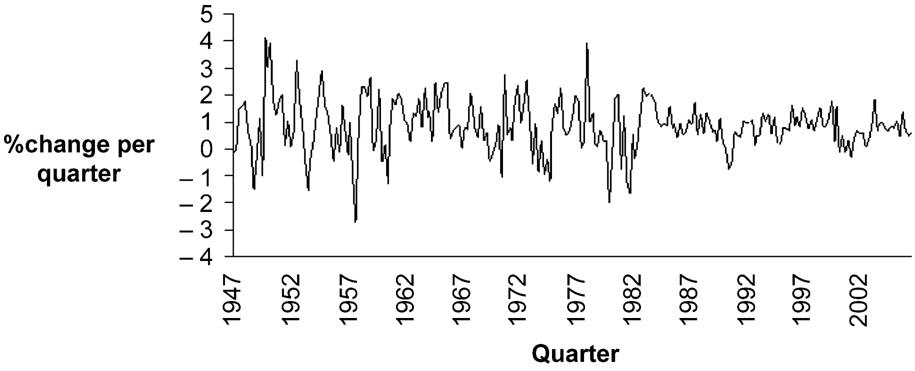
$$E_{t-1}\left[\sigma_t^2\right] = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 , \tag{3}$$

where $E_{t-1}[\,]$ is the conditional expectations operator, with expectations formed using information available in period $t-1$. Engle [46] applied this "autoregressive conditional heteroskedasticity" (ARCH) model to U.K. inflation, although in subsequent research, it has mostly been applied to financial time series. In particular, asset returns tend to display little dependence in the mean, but high positive dependence in terms of the variance (a.k.a. "volatility clustering"), which is exactly what the ARCH model was designed to capture. Beyond Engle's original paper, ARCH models have found little use in macroeconomics, although Bansal and Yaron [4] have recently attempted to resolve the so-called "equity premium puzzle" in part by assuming that US aggregate consumption growth follows a GARCH(1,1) process that generalizes Engle's original ARCH process. However, Ma [104] shows that estimates supporting a GARCH(1,1) model for aggregate consumption growth are due to weak identification, with an appropriate confidence interval suggesting little or no conditional heteroskedasticity. Weak identification is also likely a problem for the earlier application of GARCH models to macroeconomic variables by French and Sichel [49]. In general, because most macroeconomic data series are highly aggregated, the central limit theorem is relevant, at least in terms of eliminating "fat tails" due to volatility clustering that may or may not be present at the microeconomic level or at higher frequencies than macroeconomic data are typically measured.

The ARCH model begs the question of why not consider a stochastic process directly for the variance, rather than for the squares of the shocks. The short answer is a practical one. A model with "stochastic volatility" is more difficult to estimate than an ARCH model. In particular, it can be classified as a state-space model with an unobserved non-Gaussian volatility process that has a nonlinear relationship to the observable time series being modeled. In the simple case of no serial correlation in the underlying series (e. g., no AR dynamics), a stochastic volatility model can be transformed into a linear state-space model for the squares of the series, although the model still has non-Gaussian errors. However, the lack of serial correlation means that this simple version of the model would be more appropriate for applications in finance than macroeconomics. In any event, while the Kalman filter can be employed to help estimate linear Gaussian state-space models, it is less suitable for non-Gaussian state-space models and not at all suitable for nonlinear state-space models. Recent advances in computing power have made simulation-based techniques (the Gibbs sampler and the so-called "particle filter") available to estimate such models, but these techniques are far from straightforward and are highly computationally intensive. See Kim, Shephard, and Chib [88] and Chib, Nardari, and Shephard [21] on estimation of stochastic volatility models via the Gibbs sampler and particle filtering. Meanwhile, such models have rarely been applied to macroeconomic data due to the lack of interesting volatility dynamics discussed above.

To the extent that stochastic volatility models have been applied in macroeconomics, the focus has been on capturing structural change (i. e., permanent variation) in volatility rather than volatility clustering. For example, Stock and Watson [138] investigate the so-called "Great Moderation" using a stochastic volatility model and confirm the findings reported in Kim and Nelson [77] and McConnell and Perez-Quiros [107] that there was a permanent reduction in the volatility of US real GDP growth in the mid-1980s (see also [82,116,132]). This change in volatility is fairly evident in Fig. 2, which presents the time series for US real GDP growth for each quarter from 1947:Q2 to 2006:Q4.

Yet, while it is sometimes merely a matter of semantics, it should be noted that "structural change" is a dis-



Macroeconomics, Non-linear Time Series in, Figure 2
US real GDP growth 1947–2006 (Source: St. Louis Fed website)

tinct concept from "nonlinearity". In particular, *structural change* can be thought of as a change in the model describing a time series, where the change is permanent in the sense that it is not expected to be reversed. Then, if the underlying structure of each model is linear, such as for the AR(1) model in (1), there is nothing particularly "nonlinear" about structural change. On the other hand, Bayesian analysis of structural change blurs the distinction between structural change and nonlinearity. In particular, it treats parameters as random variables for the purposes of making inferences about them. Thus, the distinction between a model that allows "parameters" to change according to a stochastic process and a collection of models with the same structure, but different parameters, is essentially a matter of taste, even if the former setup is clearly nonlinear, while the latter is not. For example, consider the classic time-varying parameter model (see, for example [29]). Like the stochastic volatility model, it assumes a stochastic process for the parameters in what would, otherwise, be a linear process. Again, starting with the AR(1) model in (1) and letting $\beta = (c, \phi)'$, a time-varying parameter model typically assumes that the parameter vector evolves according to a multivariate random walk process:

$$\beta_t = \beta_{t-1} + \nu_t , \quad \nu_t \sim \text{ i.i.d.} \left(0, \Sigma\right) . \tag{4}$$

Because the time-varying parameter model treats the evolution of parameters as a stochastic process, it is clearly a nonlinear model. At the same time, its application to data provides an inherently Bayesian investigation of structural change in the relationships between dependent and independent variables, where those relationships may, in fact, be linear. In general, then, analysis of structural change in linear relationships should be considered an example of nonlinear time series analysis when nonlinear models, such as stochastic volatility models or time-varying parameter models, are used in the analysis, but structural change should not be thought of as nonlinear in itself.

In terms of macroeconomics, time-varying parameter models have recently been used to consider structural change in vector autoregressive (VAR) models of the US economy. Cogley and Sargent [26] employ such a model to argue that US inflation dynamics have changed considerably in the postwar period. Based on Sims' [135] critique that evidence for structural change in time-varying parameters may be the spurious consequence of ignoring heteroskedasticity in the error processes for a VAR model, Cogley and Sargent [27] augment their time-varying parameter model with stochastic volatility and find that their results are robust. Primiceri [123] employs a structural VAR with time-varying parameters and stochastic volatility and also finds evidence of structural changes in inflation dynamics, although he questions the role of monetary policy in driving these changes. Whether these structural changes are evident in Fig. 3, which displays US consumer price inflation for each month from 1960:M1 to 2006:M12, is debatable. However, it is fairly clear that a basic AR process with constant parameters would be an inadequate model for inflation.

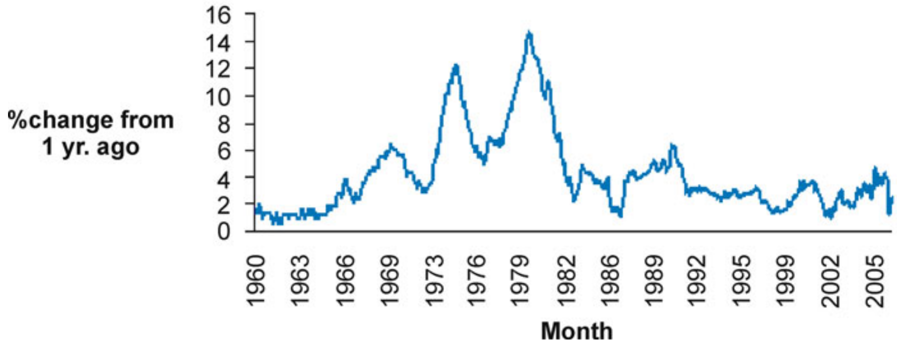
It is worth mentioning that there is a simpler time-varying parameter model that has seen considerable use in macroeconomics. It is the unobserved components (UC) model used for trend/cycle decomposition. A standard version of the model has the following form:

$$y_t = \tau_t + c_t , \tag{5}$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t , \quad \eta_t \sim \text{ i.i.d.N } \left(0, \sigma_\eta^2\right) , \tag{6}$$

$$\phi(L)c_t = \varepsilon_t , \quad \varepsilon_t \sim \text{ i.i.d.N } \left(0, \sigma_\varepsilon^2\right) , \tag{7}$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, the roots of $\phi(z) = 0$ lie outside the unit circle, and $\text{corr}\,(\eta_t, \varepsilon_t) = \rho_{\eta\varepsilon}$. It is pos-



Macroeconomics, Non-linear Time Series in, Figure 3
US inflation 1960–2006 (Source: St. Louis Fed website)

sible to think of the UC model as a time-varying parameter model in which the unconditional mean of the process is equal to the trend τ_t, meaning that it undergoes structural change, rather than remaining constant, as it does for the AR(1) process described by (1). A glance at the upward trajectory of real GDP in Fig. 1 makes it clear that a basic AR process would be an extremely bad model for the time series. Indeed, Morley, Nelson, and Zivot [114] applied the model in (5)–(7) to 100 times the natural logarithms of US real GDP under the assumption that the lag order $p = 2$ and with no restrictions on the correlation between η_t and ε_t and found that most of the variation in log real GDP was due to the trend rather than the AR cycle c_t (note that natural logarithms are more appropriate for time series modeling than the raw data in Fig. 1 because the "typical" scale of variation for real GDP is more closely linked to percentage changes than to absolute changes). Yet, while the UC model can be thought of as a time-varying parameter model, it is not, in fact, nonlinear. In particular, the UC model for log real GDP is equivalent to an autoregressive moving-average (ARMA) model for the first differences of log real GDP. Likewise, the AR(1) model in (1) may be very sensible for real GDP growth in Fig. 2, even though it would be a bad model for real GDP in Fig. 1. In general, if it is possible to transform a time series, such as going from Fig. 1 to Fig. 2, and employ a linear model for the transformed series, then the time series analysis involved is linear. Likewise, under this formulation, the simple version of the stochastic volatility model for a series with no serial correlation also falls under the purview of linear time series analysis. Only time-varying parameter and stochastic volatility models that cannot be transformed into linear representations are nonlinear.

Of course, the semantics over "linear" and "nonlinear" are hardly important on their own. What is important is whether structural change is mistaken for recurring changes in parameters or vice versa. In terms of structural VAR models for the US economy, Sims and Zha [136] argue that when parameters are allowed to undergo large, infrequent changes, rather than the smaller, more continuous changes implied by a time-varying parameter model, there is no evidence for changes in dynamic structure of postwar macroeconomic data. Instead, there are only a few large, infrequent changes in the variance of shocks. Furthermore, among the models that assume some change in dynamics, their Bayesian model comparison favors a model in which only the monetary policy rule changes. Among other things, these findings have dramatic implications for the Lucas [100,101] critique, which suggests that correlations between macroeconomic variables should be highly sensitive to changes in policy, thus leaving successful forecasting to "structural" models that capture optimizing behavior of economic agents, rather than "reduced-form" models that rely on correlations between macroeconomic structures. The results in Sims and Zha [136] suggest that the Lucas critique, while an interesting theoretical proposition with the virtue of being empirically testable, is not, in fact, supported by the data.

From the point of view of time series analysis, an interesting aspect of the Sims and Zha [136] paper and earlier papers on structural change in the US economy by Kim and Nelson [77] and McConnell and Perez-Quiros [107] is that they consider nonlinear regime-switching models that allow for changes in parameters to be recurring. That is, while the models can capture structural change, they do not impose it. Using univariate regime-switching models of US real GDP growth, Kim and Nelson [77] and McConnell and Perez-Quiros [107] find a one-time permanent reduction in output growth volatility in 1984. However, using their regime-switching VAR model, Sims and Zha [136] find that a small number of volatility regimes recur multiple times in the postwar period. In terms of the earlier discussion about the lack of volatility dynamics in macroeconomic data, this finding suggests that there are some volatility dynamics after all, but these dynamics correspond to less frequent changes than would be implied by ARCH or a continuous stochastic volatility process. More generally, the allowance for recurring regime switches is relevant because time series models with regime switches have been the most successful form of nonlinear models in macroeconomics. However, for reasons discussed in the next section, regime-switching models are typically employed to capture changing dynamics in measures of economic activity over different phases of the business cycle, rather than structural change in inflation or recurring changes in shock variances.

To summarize this section, there are different types of nonlinear time series models employed in macroeconomics. While models that assume a nonlinear specification for the relationship between observable variables exist (e. g., the bilinear model), they are rarely used in practice. By contrast, models that allow some parameters to undergo changes over time are much more common in macroeconomics. The examples discussed here are ARCH models, stochastic volatility models, time-varying parameter models, and regime-switching models. When examining structural change, there is a conceptual question of whether the analysis is "linear" or "nonlinear". However, as long as the process of structural change is an explicit part of the model (e. g., the time-varying parameter model), and excluding cases where it is possible to transform the model to have a linear representation (e. g., the UC model to an ARMA model), the analysis can be thought of as nonlinear. Meanwhile, time series analysis of recurring regime switches is unambiguously nonlinear. As discussed in the next section, nonlinear regime-switching models come in many versions and have found wide use in macroeconomics modeling business cycle asymmetry.

Business Cycle Asymmetry

The topic of business cycle asymmetry is broad and the literature on it extensive. As a result, it is useful to divide the discussion in this section into four areas: i) concepts of business cycle asymmetry and their relationships to nonlinearity; ii) nonlinear models of business cycle asymmetry; iii) evidence for nonlinear forms of business cycle asymmetry; and iv) the relevance of nonlinear forms of business cycle asymmetry for macroeconomics.

Concepts

Notions of business cycle asymmetry have a long tradition in macroeconomics. Classic references to the idea that recessions are shorter, sharper, and generally more volatile than expansions are Mitchell [109], Keynes [72], and Burns and Mitchell [13]. For example, in his characteristic style, John Maynard Keynes writes, "…the substitution of a downward for an upward tendency often takes place suddenly and violently, whereas there is, as a rule, no such sharp turning point when an upward is substituted for a downward tendency." (see p. 314 in [72]). Similarly, albeit more tersely, Wesley Mitchell writes, "…the most violent declines exceed the most considerable advances. The abrupt declines usually occur in crises; the greatest gains occur in periods of revival… Business contractions appear to be a briefer and more violent process than business expansions." (see p. 290 in [109]). Milton Friedman also saw business cycle asymmetry in the form of a strong relationship between the depth of recession and the strength of a recovery, with no corresponding relationship between the strength of an expansion with the severity of the subsequent recession (see [50,51]).

The link between business cycle asymmetry and nonlinearity depends, in part, on the definition of "business cycle". Harding and Pagan [67] discuss three possible definitions that are presented here using slightly modified terminology. Based on the work of Burns and Mitchell [13], the first definition is that the business cycle is the alternation between phases of expansion and recession in the *level* of economic activity. The second definition, which is often left implicit when considered, is that the business cycle represents transitory *deviations* in economic activity from a permanent or "trend" level. The third definition, which is also often only implicitly considered, is that the business cycle corresponds to any short-run *fluctuations* in economic activity, regardless of whether they are permanent or transitory.

Under the "level" definition of the business cycle, there is nothing inherently nonlinear about asymmetry in terms of the duration of expansions and recessions. Positive drift in the level of economic activity implies longer expansions than recessions, even if the underlying process is linear. Even asymmetry in the form of relative sharpness and steepness of a recession alluded to in the above quote from Keynes does not necessarily indicate nonlinearity. Again, given positive drift, an outright decline in economic activity only occurs when there are large negative shocks to the underlying process, while an expansion occurs for all positive shocks and small negative shocks. Thus, a recession is likely to look like a relatively sharp reversal in the level. Furthermore, with positive serial correlation in growth, such as implied by a linear AR(1) process as in (1) with $\phi > 0$, recessions will appear steeper than expansions due to the dynamic effects of large negative shocks. On the other hand, as discussed in more detail later, nonlinear models are much more successful than linear models at reproducing business cycle asymmetry in the form of a strong link between recessions and their recoveries versus a weak link between expansions and subsequent recessions noted by Friedman [50].

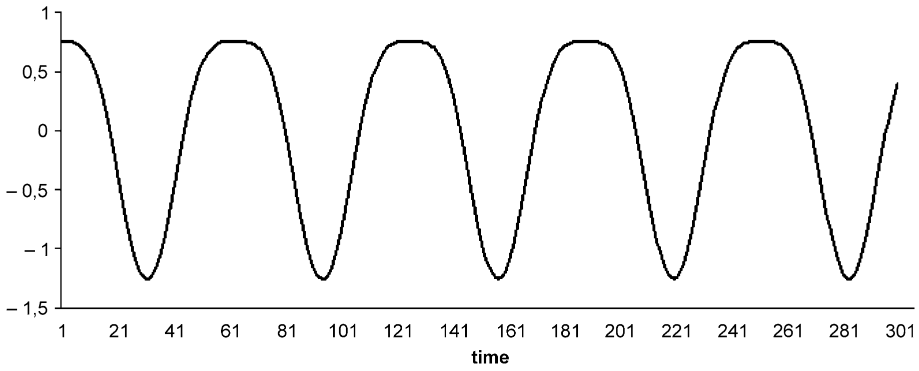
Under the "deviations" definition of the business cycle, asymmetry is closely linked to nonlinearity. While it is possible for asymmetry in the independent and identical distribution of the underlying shocks to generate asymmetry in a linear process, any persistence in the process would severely dampen the asymmetries in the unconditional distribution. Thus, under the assumption that the transitory component of economic activity is at least somewhat persistent, asymmetries such as differences in the durations of positive and negative deviations from trend or relative sharpness and steepness in negative deviations compared to positive deviations are more suggestive of nonlinear dynamics (i. e., changing correlations) than underlying asymmetric shocks.

Under the "fluctuations" definition of the business cycle, the link between nonlinearity and asymmetry also depends on the relative roles of shocks and dynamics in generating asymmetries. However, because growth rates are less persistent than most measures of the transitory component of economic activity and because they mix together permanent and transitory shocks that may have different means and variances, it is quite plausible that asymmetry in the distribution of shocks is responsible for asymmetry in growth rates. Of course, nonlinear dynamics are also a plausible source of asymmetry for growth rates.

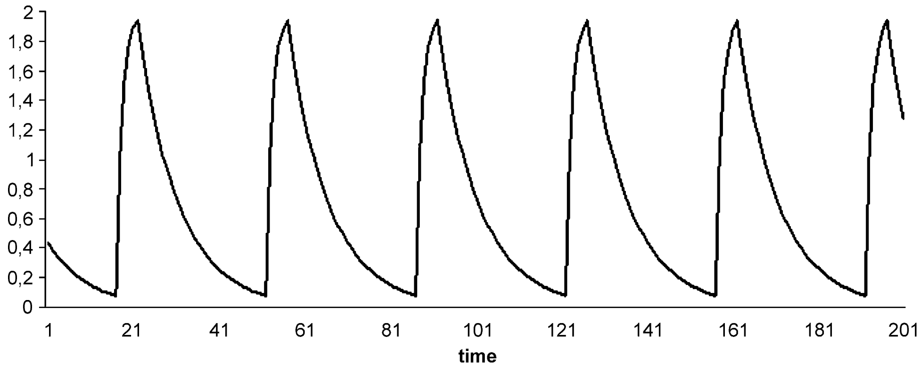
In terms of asymmetries, it is useful to consider the formal classifications developed and discussed in Sichel [133], McQueen and Thorley [108], Ramsey and Rothman [124], Clements and Krolzig [24], and Korenok, Mizrach, and Radchenko [95] of "deepness", "steepness", and "sharpness". Following Sichel [133], a process is said to have *deepness* if its unconditional distribution is skewed and *steepness* if the distribution of its first-differences is skewed. Following McQueen and Thorley [108], a process is said to have *sharpness* if the probability of a peak occurring when it has been increasing is different than the probability of a trough occurring when it has been decreasing. However, despite these definitions, the different types of asymmetries are most easily understood with visual examples.

Figure 4 presents an example of a simulated time series with deepness, with the distance from peak of the cycle to the mean less than the distance from the mean to trough of the cycle (see [124], for the details of the process generating this time series). In addition to deepness, the series appears to display sharpness in recessions, with the peak of the cycle more rounded than the trough, although the fact that the simulated series is deterministic means it cannot be directly related to the definition of sharpness in McQueen and Thorley [108] mentioned above. Meanwhile, there is no steepness because the slope from peak to trough is the same magnitude as the slope from trough to peak.

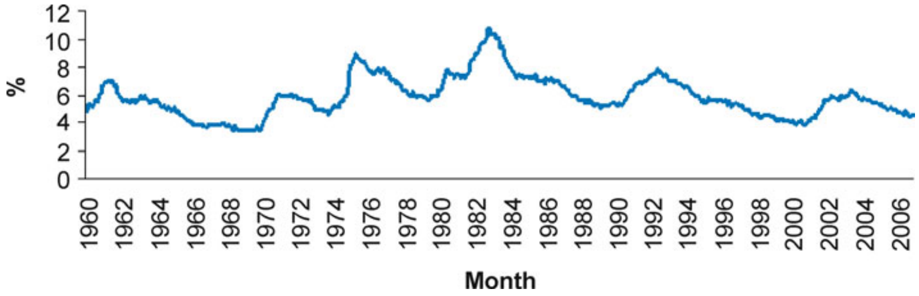
As discussed in Ramsey and Rothman [124], these different types of asymmetry can be classified in two broader categories of "time reversible" and "time irreversible". *Time reversibility* means that the substitution of $-t$ for t in the equations of motion for a process leaves the process unchanged. The upward drift that is present in many macroeconomic time series (such as real GDP) is clearly time irreversible. More generally, the issue of time re-



Macroeconomics, Non-linear Time Series in, Figure 4
A "deep" cycle (Source: Author's calculations based on Ramsey and Rothman [124])



Macroeconomics, Non-linear Time Series in, Figure 5
A "steep" cycle (Source: Author's calculations)



Macroeconomics, Non-linear Time Series in, Figure 6
US civilian unemployment rate 1960–2006 (Source: St. Louis Fed website)

versibility is relevant for determining whether business cycle asymmetry corresponds to deepness and sharpness, which are time reversible, or steepness, which is time irreversible. For example, the time series in Fig. 4 can be flipped on the vertical axis without any resulting change. Thus, it is time reversible. By contrast, consider the simulated time series with "steepness" in Fig. 5. The series is generated from a regime-switching process with asymmetric shocks across two regimes and different persistence for shocks in each regime. In this case, flipping the series on the vertical axis would produce flat inclines and steep declines. Thus, it is time irreversible.

The relevance of the distinction between time reversible and time irreversible processes is obvious from Fig. 6, which presents the time series for the US civilian unemployment rate for each month from 1960:M1 to 2006:M12. The inclines are steep relative to the declines. Thus, there is a clear visual suggestion of the steepness form of asymmetry. Indeed, the modern literature on business cycle asymmetry begins with Neftçi's [115] investigation of this issue using a nonlinear regime-switching model in which the prevailing "business cycle" regime in a given period is assumed to depend on a discrete Markov process driven by whether the US unemployment rate is rising or falling in that period. Given the link to the first differences of the unemployment rate, his finding that the continuation probabilities for the two regimes are different, with declines more likely to persist than increases, provides formal support for the presence of the steepness forms of asymmetry in the unemployment rate (also, see [127]). It should also be noted that, while not related to time irreversibility, the different continuation probabilities also directly imply sharpness.

Models

The subsequent literature on regime-switching models in macroeconomics can be usefully divided into two categories that are both related to Neftçi's [115] model. First, *Markov-switching models* assume that the prevailing regime depends on an unobserved discrete Markov process. The main distinction from Neftçi [115] is that the Markov process is unobserved (hence, these models are sometimes referred to as a "hidden Markov models"). Second, *self-exciting threshold models* assume that the prevailing regime is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values, much like the regime in Neftçi's [115] model depends on whether the change in the unemployment rate was positive or negative.

Hamilton [59] is the seminal paper in terms of Markov-switching models. His model has a basic AR structure, like in (1), but for the first-differences of the time series of interest:

$$\phi(L)\left(\Delta y_t - \mu_t\right) = \varepsilon_t , \quad \varepsilon_t \sim \text{i.i.d.}\,(0, \sigma^2) \,, \tag{8}$$

where Δy_t is 100 times the change in the natural logarithm of real Gross National Product (GNP). The only difference from a linear AR model is that the mean follows a stochastic process:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) \,, \tag{9}$$

with the indicator function $I(S_t = j)$ equal to 1 if $S_t = j$ and 0 otherwise and $S_t = \{1, 2\}$ following an unobserved discrete Markov state variable that evolves according to the following fixed transition matrix:

$$\begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} ,$$

where $p_{ij} \equiv \Pr[S_t = j | S_{t-1} = i]$ and the columns sum to one.

There are two aspects of Hamilton's [59] model that should be mentioned. First, while the demeaned specification is equivalent to a regression model specification (e. g., (1)) in the linear setting, with $\mu = c/(1 - \phi)$, the two specifications are no longer equivalent in the nonlinear setting. In particular, if the intercept c were switching instead of the mean μ, then past regime switches would be propagated by the AR dynamics (see [61], for an example of such a model). By contrast, with μ switching, there is a clear separation between the "nonlinear" dynamics due to the evolution of the state variable (which does alter the correlations between Δy_t and its lags) and the "linear" dynamics due to the ε_t shocks and the AR parameters. Second, in order to eliminate arbitrariness in the labeling of states, it is necessary to impose a restriction such as $\mu_1 > \mu_2$, which corresponds to higher mean growth in state 1 than in state 2. Furthermore, given the application to output growth, if $\mu_1 > 0$ and $\mu_2 < 0$, the states 1 and 2 can be labeled "expansion" and "recession", respectively.

Hamilton's [59] paper had a big impact on econometrics and macroeconomics for two reasons. First, it included an elegant filter that could be used to help estimate Markov-switching models via maximum likelihood and, along with a smoother, calculate the posterior distribution of the unobserved state variable (filters and smoothers are recursive algorithms that make inferences about unobserved state variables, with filters considering only information available at the time the state variable is realized and smoothers incorporating any subsequent available information). Second, the resulting posterior probability of the "recession" regime corresponded closely to the National Bureau of Economic Research (NBER) dating of recessions. The NBER dating is based on non-structural and subjective analysis of a wide variety of indicators. The official line from its website is "The NBER does not define a recession in terms of two consecutive quarters of decline in real GDP. Rather, a recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales." (www.nber.org/cycles/cyclesmain.html). Thus, it is, perhaps, remarkable that a simple time series model using only information in real GNP could find such similar dates for recessions. Of course, as emphasized by Harding and Pagan [66], a simple rule like "two consecutive quarters of decline in real GDP" also does extremely well in matching the NBER recessions, regardless of NBER claims that it is not following such a rule. Yet, more important is the notion implied by Hamilton's [59] results that the NBER is identifying a meaningful structure in the economy, rather than simply reporting (sometimes with considerable lag) that the economy had an episode of prolonged negative growth. Specifically, "recession" appears to be an indicator of a different state for the dynamics of the economy, rather than a label for particular realizations of linear process. (As an aside, the fact that the popular press pays so much attention to NBER pronouncements on recessions also supports the idea that it is identifying a meaningful macroeconomic structure).

Numerous modifications and extensions of Hamilton's [59] model have been applied to macroeconomic data. For example, while estimates for Hamilton's [59] model imply that the linear ε_t shocks have large permanent effects on the level of real GDP, Lam [96] considers a model in which the only permanent shocks to real GNP are due to regime switches. Despite this very different assumption, he also finds that the regime probabilities implied by his model correspond closely to NBER dating of expansions and recessions. Kim [74] develops a filter that can be used for maximum likelihood estimation of state-space models with Markov-switching parameters and confirms the results for Lam's [96] model. Motivated by Diebold and Rudebusch's [38] application of Hamilton's [59] model to the Commerce Department's coincident index of economic activity instead of measures of aggregate output such as real GNP or real GDP, Chauvet [19] employs Kim's [74] filter to estimate an unobserved components model of a coincident index using Hamilton's [59] model as the specification for its first differences. Other multivariate extensions include Kim and Yoo [87], Ravn and Sola [125], Kim and Nelson [76], Kim and Murray [75], Kim and Piger [81], Leamer and Potter [97], Camacho [14], and Kim, Piger, and Startz [84]. The general theme of these studies is that the multivariate information, such as coincident indicators or aggregate consumption and investment, helps to strongly identify the nonlinearity in economic activity, with regimes corresponding even more closely to NBER dates than for univariate analysis based on real GNP or real GDP.

In terms of business cycle asymmetry, an important extension of Hamilton's [59] model involves allowing for three regimes to capture three phases of the business cycle: "expansion", "recession", and "recovery" (see [134]). Papers with three-regime models include Boldin [8], Clements and Krolzig [23], and Leyton and Smith [98]. The specification in Boldin [8] modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \mu_3 \cdot I(S_t = 3)\,, \quad (10)$$

where $S_t = \{1, 2, 3\}$ has fixed transition matrix:

$$\begin{bmatrix} p_{11} & 0 & p_{31} \\ p_{12} & p_{22} & 0 \\ 0 & p_{23} & p_{33} \end{bmatrix} .$$

The zeros in the transition matrix restrict the state sequence to follow the pattern of $\{S_t\} = \cdots 1 \to 2 \to 3 \to 1 \cdots$. Given the normalization $\mu_1 > \mu_2$, the restriction on the transitional matrix implies that the economy goes from expansion to recession to recovery and back to expansion. While there is no restriction on μ_3, Boldin [8] finds it is greater than μ_1, which means that the third regime corresponds to a high-growth recovery. As discussed in Clements and Krolzig [24], this third regime allows for steepness in output growth, while the basic two-regime Hamilton [59] model can only capture deepness and sharpness (the two are inextricably linked for a two-regime model) in growth. Note, however, from the definitions presented earlier, deepness in growth implies steepness the level of output.

It is possible to capture high-growth recoveries without resorting to three regimes. For example, Kim and Nelson [79] develop an unobserved components model that assumes two regimes in the transitory component of US real GDP. A slightly simplified version of their model is given as follows:

$$y_t = \tau_t + c_t\,, \quad (11)$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t, \quad \eta_t \sim \text{i.i.d.N}\left(0, \sigma_\eta^2\right)\,, \quad (12)$$

$$\phi(L)c_t = \lambda \cdot I(S_t = 2) + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.N}\left(0, \sigma_\varepsilon^2\right)\,, \quad (13)$$

where y_t is 100 times log real GDP, $S_t = \{1, 2\}$ is specified as in Hamilton's [59] model, and state 2 is identified as the recession regime by the restriction $\lambda < 0$ (see [112,113], on the need for and implications of this restriction). Unlike Morley, Nelson, and Zivot [114], Kim and Nelson [79] impose the restriction that $\rho_{\eta\varepsilon} = 0$ in estimation, which they conduct via approximate maximum likelihood using the Kim [74] filter. As with Hamilton [59] and Lam [96], the regimes correspond closely to NBER-dated expansions and recessions. However, because the regime switching is in the transitory component only, the transition from state 1 to state 2 corresponds to a downward "pluck" in economic activity that is followed by a full recovery to trend after the transition from state 2 to state 1. Kim and Nelson [79] motivate their model as nesting Friedman's [50,51] plucking model, which assumes output cannot exceed a ceiling level, but is occasionally plucked below full capacity by recessionary shocks resulting from activist monetary policy. In line with Friedman's observations, Kim and Nelson's [79] model relates the strength of a recovery to the severity of the preceding recession, with no corresponding link between the strength of an expansion and the severity of a recession (see also [2,134,150]). Notably, the transitory component for their estimated model achieves the trifecta of business cycle asymmetries in the form of deepness, steepness, and sharpness.

Another model that captures three phases of the business cycle with only two underlying regimes is the "bounceback" model of Kim, Morley, and Piger [83]. The model modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I\left(S_{t-j} = 2\right), \quad (14)$$

where the number of lagged regimes to consider in the third term on the right hand side of (14) is determined by the discrete "memory" parameter m, which is estimated to be six quarters for US postwar quarterly real GDP. Given the restriction $\mu_1 > \mu_2$, the third term can be interpreted as a pressure variable that builds up the longer a recession persists (up to m periods, where $m = 6$ quarters is long enough to capture all postwar recessions) and is motivated by the "current depth of recession" variable of Beaudry and Koop [6] discussed later. Then, if $\lambda > 0$, growth will be above μ_1 for up to the first six quarters of an expansion. That is, there is a post-recession "bounceback" effect, as in Kim and Nelson's [79] plucking model. Meanwhile, the specification in (14) can be thought of as a "u-shaped recession" version of the model because the pressure variable starts mitigating the effects of a recession the longer the regime persists. Morley and Piger [111] consider a slightly modified "v-shaped recession" version of the model that assumes the pressure variable only affects growth after the recession ends, thus producing a sharper trough:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I(S_t = 1) \cdot I\left(S_{t-j} = 2\right) . \quad (15)$$

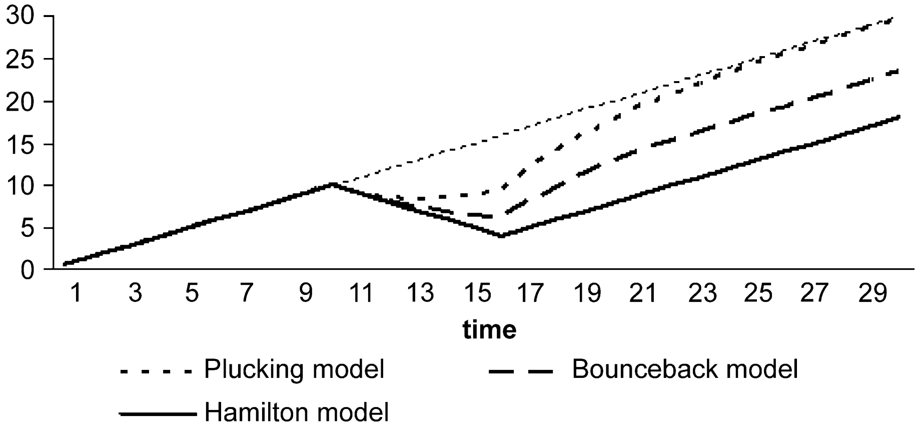
This version of the model is identical to Hamilton's [59] model in all but the first m periods of an expansion. Finally, Morley and Piger [113] consider a "depth" version of the model that relates the pressure variable to both the length and severity of a recession:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} \left(\mu_1 - \mu_2 - \Delta y_{t-j}\right) \cdot I\left(S_{t-j} = 2\right) . \quad (16)$$

In this case, the post-recession bounceback effect depends on the relative severity of a recession. Regardless of the specification, the estimated bounceback effect for US real GDP based on maximum likelihood estimation via the Hamilton [59] filter is large (see [83,111,113]).

While Kim, Morley, and Piger's [83] bounceback model can capture "plucking" dynamics, there is no restriction that regime switches have only transitory effects. Instead, the model nests both the Hamilton [59] model assumption that recessions have large permanent effects in the case that $\lambda = 0$ and Kim and Nelson's [79] "plucking" model assumption that recessions have no permanent effects in the case that $\lambda = (\mu_1 - \mu_2)/m$ (for the specification in (14)). Figure 7 presents examples of simulated time series for the plucking model, the bounceback model, and the Hamilton model. In each case, "output" is subject to a recession regime that lasts for six periods. For the plucking model, output returns to the level it would have been in the absence of the recession. For the Hamilton model, output is permanently lower as a result of the recession. For the bounceback model, recessions can have permanent effects, but they will be less than for the Hamilton model if $\lambda > 0$ (indeed, if $\lambda > (\mu_1 - \mu_2)/m$, the long-run path of the economy can be increased by recessions, a notion related to the "creative destruction" hypothesis of Schumpeter [131]). In practice, Kim, Morley, and Piger [83] find a very small negative long-run impact of US recessions, providing support for the plucking model dynamics and implying considerably lower economic costs of recessions than the Hamilton model.

Another extension of Hamilton's [59] model involves relaxing the assumption that the transition probabilities for the unobserved state variable are fixed over time (see [39]). Filardo [48] considers time-varying transition probabilities for a regime-switching model of industrial production growth where the transition probabilities depend on leading indicators of economic activity. Durland and McCurdy [40] allow the transition probabilities for



Macroeconomics, Non-linear Time Series in, Figure 7
Simulated paths for "Output" (Source: Author's calculations)

real GNP growth to depend on the duration of the prevailing regime. DeJong, Liesenfeld, and Richard [34] allow the transition probabilities for real GDP growth depend on an observed "tension index" that is determined by the difference between recent growth and a "sustainable" rate that corresponds to growth in potential output. Kim, Piger, and Startz [84] allow for a dynamic relationship between multiple unobserved discrete state variables in a multivariate setting and find that regime-switches in the permanent component of economic activity tend to lead regime-switches in the transitory component when the economy heads into recessions.

The distinction between Markov-switching models and threshold models is blurred somewhat by time-varying transition probabilities. A standard demarcation is that Markov-switching models typically assume the discrete state variables driving changes in regimes are exogenous, while threshold models allow for endogenous switching. However, this exogenous/endogenous demarcation is less useful than it may at first appear. First, as is always the problem in macroeconomics, it is unlikely that the variables affecting time-varying transition probabilities are actually strictly exogenous, even if they are predetermined. Second, Kim, Piger and Startz [85] have developed an approach for maximum likelihood estimation of Markov-switching models that explicitly allow for endogenous switching. In terms of macroeconomics, Sinclair [137] applies their approach to estimate a version of the regime-switching UC model in (11)–(13) for US real GDP that allows for a non-zero correlation between the regular shocks η_t and ε_t, as in Morley, Nelson, and Zivot [114], as well as dependence between these shocks and the unobserved state variable S_t that generates downward plucks in output. She finds that permanent shocks are more important than suggested by Kim and Nelson's [79] estimates. However, she confirms the importance of the plucking dynamic, with a test supporting the standard exogeneity assumption for the discrete Markov-switching state variable.

Another demarcation that would seem to provide a possible means of distinguishing between Markov-switching models and threshold models arises from the fact that, starting from an AR specification, threshold models typically extend the basic model by allowing for changes in AR parameters, while, as discussed earlier, Markov-switching models typically extend the model by allowing for changes in the mean. However, this demarcation is also less useful than it may at first appear since Markov-switching models have alternative representations as autoregressive processes (see [59]). Furthermore, some threshold models assume constant AR parameters (e. g., [120]). In particular, regardless of presentation, both types of models capture nonlinear dynamics in the conditional mean.

The more general and useful demarcation between Markov-switching models and threshold models is that the prevailing regime is unobservable in the former, while it is observed in the latter. Meanwhile, the observable regimes in threshold models make it feasible to consider more complicated transitions between regimes than Markov switching models. In particular, it is possible with a threshold model to allow a mixture of regimes to prevail in a given time period.

Tong [145] introduced the basic threshold autoregressive (TAR) model. In a "self-exciting" TAR model, the autoregressive coefficient depends on lagged values of the time series. For example, a simple two-regime AR(1) TAR model is given as follows:

$$\begin{aligned} y_t &= c + \phi^{(1)} \cdot I\left(y_{t-m} < \tau\right) \cdot y_{t-1} \\ &+ \phi^{(2)} \cdot I\left(y_{t-m} \geqslant \tau\right) \cdot y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2), \end{aligned} \tag{17}$$

where $\phi^{(1)}$ and $\phi^{(2)}$ are the AR(1) parameters associated with the two regimes, τ is the threshold, and m is the discrete delay parameter. A variant of the basic TAR model that allows multiple regimes to prevail to different degrees is the smooth transition autoregressive (STAR) model (see [18,58,140,142]). For STAR models, the indicator function is replaced by transition functions bounded between zero and one. The STAR model corresponding to (17) is

$$\begin{aligned} y_t = c &+ \phi^{(1)} \cdot F_1\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} \\ &+ \phi^{(2)} \cdot F_2\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} + \varepsilon_t, \\ &\varepsilon_t \sim \text{i.i.d.}\left(0, \sigma^2\right), \end{aligned} \tag{18}$$

where $F_2(y_{t-m}|\tau, \gamma) = 1 - F_1(y_{t-m}|\tau, \gamma)$ and γ is a parameter that determines the shape of the transition function (in general, the larger γ, the closer the STAR model is to the TAR model). The two most popular transition functions are exponential (ESTAR) and logistic (LSTAR). The exponential transition function is

$$F_1^e = 1 - \exp\left(-\gamma(y_{t-m} - \tau)^2\right), \quad \gamma > 0, \tag{19}$$

while the logistic transition function is

$$F_1^l = \left[1 + \exp\left(-\gamma(y_{t-m} - \tau)\right)\right]^{-1}, \quad \gamma > 0. \tag{20}$$

For STAR models the transition functions are such that the prevailing autoregressive dynamics are based on a weighted average of the autoregressive parameters for each regime, rather than reflecting only one or the other, as in TAR models.

In terms of macroeconomics, both TAR and STAR models have been employed to capture business cycle asymmetry. A key question is what observed threshold might be relevant. On this issue, a highly influential paper is Beaudry and Koop [6]. Related to the notion discussed above that recessions represent a meaningful macroeconomic structure, they consider whether real GDP falls below a threshold defined by its historical maximum. Specifically, they define a "current depth of recession" (CDR) variable as follows:

$$\mathrm{CDR}_t = \max\left\{y_{t-j}\right\}_{j\geqslant 0} - y_t\ . \tag{21}$$

Figure 8 presents the current depth of recession using US real GDP for each quarter from 1947:Q1 to 2006:Q4.

Beaudry and Koop [6] augment a basic linear ARMA model of US real GNP growth with lags of the CDR variable. They find that the inclusion of the CDR variable implies much less persistence for large negative shocks than for small negative shocks or positive shocks. The asymmetry in terms of the response of the economy to shocks corresponds closely to the idea discussed earlier that deep recessions produce strong recoveries. Indeed, the Beaudry and Koop [6] paper provided a major motivation for most of the extensions of Hamilton's [59] model discussed earlier that allow for high-growth recoveries.

In terms of threshold models in macroeconomics, Beaudry and Koop [6] initiated a large literature. Tiao and Tsay [144], Potter [121], and Clements and Krolzig [23] consider two-regime TAR models with the threshold either fixed at zero or estimated to be close to zero. Pesaran and Potter [120] and Koop and Potter [91] consider a three-regime TAR model (with many restrictions for tractability) that incorporates the CDR variable and an "overheating" (OH) variable reflecting cumulated growth following large positive shocks. Specifically, a simple homoskedastic, AR(1) version of Pesaran and Potter's [120] "floor and ceiling" model is given as follows:

$$\begin{aligned}\Delta y_t &= c + \phi\Delta y_{t-1} + \lambda_1\mathrm{CDR}_{t-1} + \lambda_2\mathrm{OH}_{t-1} + \varepsilon_t\,,\\ &\qquad \varepsilon_t \sim N(0,\sigma^2)\,,\end{aligned} \tag{22}$$

where

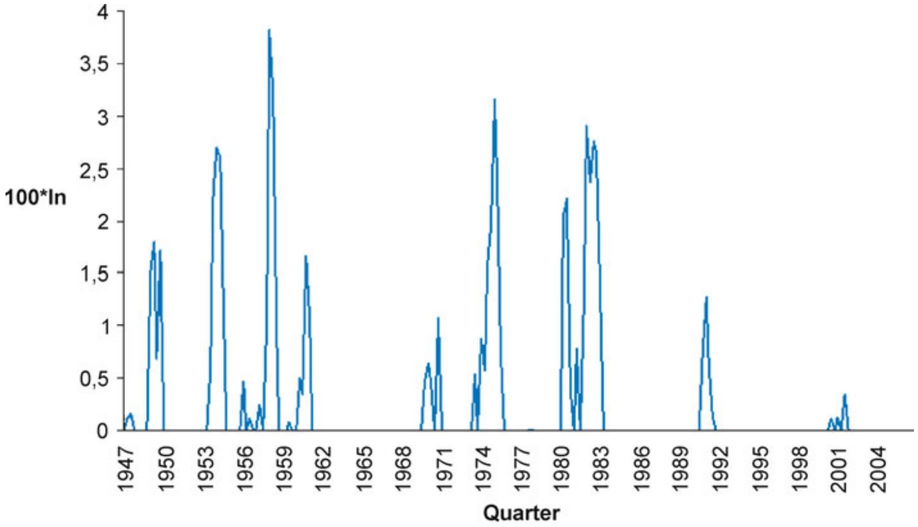
$$\begin{aligned}\mathrm{CDR}_t &= -(\Delta y_t - \tau_F)\cdot F_t\cdot(1-F_{t-1})\\ &\quad + (\mathrm{CDR}_{t-1} - \Delta y_t)\cdot F_t\cdot F_{t-1}\,,\end{aligned} \tag{23}$$

$$\begin{aligned}F_t &= I\left(\Delta y_t < \tau_F\right)\cdot(1-F_{t-1})\\ &\quad + I\left(\mathrm{CDR}_{t-1} - \Delta y_t > 0\right)\cdot F_{t-1}\,,\end{aligned} \tag{24}$$

$$\mathrm{OH}_t = (\mathrm{OH}_{t-1} + \Delta y_t - \tau_C)\cdot C_t\,, \tag{25}$$

$$C_t = (1-F_t)\cdot I\left(\Delta y_t > \tau_C\right)\cdot I\left(\Delta y_{t-1} > \tau_C\right). \tag{26}$$

The indicator variable $F_t = \{0,1\}$ denotes whether the economy is in the "floor" regime, while $C_t = \{0,1\}$ denotes whether the economy is in the "ceiling" regime. The



Macroeconomics, Non-linear Time Series in, Figure 8
Current depth of recession 1947–2006 (Source: Author's calculations based on Beaudry and Koop [6])

CDR variable is the same as in (20) if the threshold $\tau_F = 0$. Thus, a high-growth post-recession recovery is implied by $\lambda_1 > 0$. In particular, with $\tau_F = 0$, the "floor" regime is activated when real GDP falls below its historical maximum at the onset of a recession and remains activated until output recovers back to its pre-recession level. The OH variable captures whether real GDP is above a sustainable level based on the threshold level τ_C of growth. A capacity-constraint effect is implied by $\lambda_2 < 0$. Note, however, that the "ceiling" regime that underlies the OH variable can be activated only when the "floor" regime is off, ruling out the possibility that a high-growth recovery from the trough of a recession is labeled as "overheating". There is also a requirement of two consecutive quarters of fast growth above the threshold level τ_C in order to avoid labeling a single quarter of fast growth as "overheating". Meanwhile, a heteroskedastic version of the model allows the variance of the shocks to evolve as follows:

$$\sigma_t^2 = \sigma_1^2 \cdot I(F_{t-1} + C_{t-1} = 0) + \sigma_2^2 F_{t-1} + \sigma_3^2 C_{t-1} \,. \quad (27)$$

Also, in a triumph of controlled complexity, Koop and Potter [92] develop a multivariate version of this model, discussed later.

A related literature on STAR models of business cycle asymmetry includes Teräsvirta and Anderson [143], Teräsvirta [141], van Dijk and Franses [148], and Öcal and Osborn [117]. Similar to the development of Markov-switching models and TAR models, van Dijk and Franses [148] develop a multi-regime STAR model and find evidence for more than two regimes in economic activity. Likewise, using U.K. data on industrial production, Öcal and Osborn [117] find evidence for three regimes corresponding to recessions, normal growth, and high growth. Rothman, van Dijk, and Franses [130] develop a multivariate STAR model to examine nonlinearities in the relationship between money and output.

While there are many different nonlinear models of economic activity, it should be noted that, in a general sense, Markov-switching models and threshold models are close substitutes for each other in terms of their abilities to forecast (see [23]) and their abilities to capture business cycle asymmetries such as deepness, steepness, and sharpness (see [24]). On the other hand, specific models are particularly useful for capturing specific asymmetries and, as discussed next, for testing the presence of nonlinear dynamics in macroeconomic time series.

Evidence

While estimates for regime-switching models often imply the presence of business cycle asymmetries, it must be acknowledged that the estimates may be more the consequence of the flexibility of nonlinear models in fitting the data than any underlying nonlinear dynamics. In the regime-switching model context, an extreme example of over-fitting comes from a basic i.i.d. mixture model. If the mean and variance are allowed to be different across regimes, the sample likelihood will approach infinity as the estimated variance approaches zero in a regime for which the estimated mean is equal to a sample observation. (It should be noted, however, that the highest local maximum of the sample likelihood for this model produces consistent estimates of the model parameters. See [73]). Thus, it is wise to be skeptical of estimates from nonlinear models and to seek out a correct sense of their precision. Having said this, the case for nonlinear dynamics that correspond to business cycle asymmetries is much stronger than it is often made out to be, although it would be a mistake to claim the issue is settled.

From the classical perspective, the formal problem of testing for nonlinearity with regime-switching models is that the models involve *nuisance parameters* that are not identified under the null hypothesis of linearity, but influence the distributions of test statistics. For example, Hamilton's [59] model outlined in (8)–(9) collapses to a linear AR model if $\mu_1 = \mu_2$. However, under this null hypothesis, the two independent transition probabilities p_{11} and p_{22} in the transition matrix will no longer be identified (i. e., they can take on different values without changing the fit of the model). The lack of identification of these nuisance parameters is referred to as the Davies [32] problem and it means that test statistics of the null hypothesis such as a t-statistic or a likelihood ratio (LR) statistic will not have their standard distributions, even asymptotically. An additional problem for Markov-switching models is that the null hypothesis of linearity often corresponds to a local maximum for the likelihood, meaning that the score is identically zero for some parameters, thus violating a standard assumption in classical testing theory. The problem of an identically zero score is easily seen by noting that one of the fundamental tests in classical statistics, the Lagrange multiplier (LM) test, is based on determining whether the score is significantly different than zero when imposing the null hypothesis in a more general model. For Hamilton's [59] model, the scores are zero for $\mu_d = \mu_2 - \mu_1$, p_{11}, and p_{22}. Again, identically zero scores imply nonstandard distributions for a t-statistic or an LR statistic. In practice, these nonstandard distributions mean that, if researchers were to apply standard critical values, they would over-reject linearity.

Hansen [61] derives a bound for the asymptotic distribution of a likelihood ratio statistic in the setting of unidentified nuisance parameters and identically zero scores. The bound is application-specific as it depends on the covariance function of an empirical process associated with the likelihood surface in a given setting (i. e., it is model and data dependent). The distribution of the empirical process can be obtained via simulation. In his application, Hansen [61] tests linearity in US real GNP using Hamilton's [59] model. His upper bound for the p-value of the likelihood ratio test statistic is far higher than conventional levels of significance. Thus, he is unable to reject linearity with Hamilton's [59] model. However, when he proposes an extended version of the model that assumes switching in the intercept and AR coefficients, rather than the mean as in (8)–(9), he is able to reject linearity with an upper bound for the p-value of 0.02.

In a subsequent paper, Hansen [62] develops a different method for testing in the presence of unidentified nuisance parameters that yields an exact critical value rather than an upper bound for a p-value. Again, the method requires simulation, as the critical value is model and data dependent. However, this approach assumes nonzero scores and is, therefore, more appropriate for testing threshold models than Markov-switching models. In his application for this approach, Hansen [62] tests linearity in US real GNP using Potter's [121] TAR model mentioned earlier (see also [17,63,146,147], on testing TAR models and [140], on testing STAR models). Referring back to the TAR model in (17), the threshold τ and the delay parameter m are unidentified nuisance parameters under the null of linearity (i. e., the case where the AR parameters and any other parameters that are allowed to switch in the model are actually the same across regimes). Hansen [62] finds that the p-values for a variety of test statistics are above conventional levels of significance, although the p-value for the supLM (i. e., the largest LM statistic for different values of the nuisance parameters) under the hypothesis of homoskedastic errors is 0.04, thus providing some support for nonlinearity.

Garcia [53] reformulates the problem of testing for Markov-switching considered in Hansen [61] by proceeding as if the score with respect to the change in Markov-switching parameters (e. g., $\mu_d = \mu_2 - \mu_1$ for Hamilton's [59], model) is not identically zero and examining whether the resulting asymptotic distribution for a likelihood ratio test statistic is approximately correct. The big advantage of this approach over Hansen [61] is that the distribution is no longer sample-dependent, although it is still model-dependent. Also, it yields an exact critical value instead of an upper bound for the p-value. Garcia [53] reports asymptotic critical values for some basic Markov-switching models with either no linear dynamics or a mild degree of AR(1) linear dynamics ($\phi = 0.337$) and compares these to critical values based on a simulated distribution of the LR statistic under the null of linearity and a sample size of 100. He finds that his asymptotic critical values are similar to the simulated critical values for the simple models, suggesting that they may be approximately correct despite the problem of an identically zero score. The asymptotic critical values are considerably smaller than the simulated critical values in the case of Hamilton's [59] model with an AR(4) specification, although this is perhaps due to small sample issues rather than approximation error for the asymptotic distribution. Regardless, even with the asymptotic critical values, Garcia [53] is unable to reject linearity for US real GNP using Hamilton's [59] model at standard levels of significance, although the p-value is around 0.3 instead of the upper bound of around 0.7 for Hansen [61].

It is worth mentioning that the simulated critical values in Garcia's [53] study depend on the values of parameters used to simulate data under the null hypothesis. That is, the LR statistic is not *pivotal*. Thus, the approach of using the simulated critical values to test linearity would correspond to a parametric bootstrap test (see [105,106], for excellent overviews of bootstrap methods). The use of bootstrap tests (sometimes referred to as Monte Carlo tests, although see MacKinnon [105,106], for the distinction) for Markov-switching models has been limited (although see [96], for an early example) for a couple of reasons. First, the local maximum at the null hypothesis that is so problematic for asymptotic theory is also problematic for estimation. While a researcher is likely to re-estimate a nonlinear model using different starting values for the parameters when an optimization routine converges to this or another local maximum in an application, it is harder to do an exhaustive search for the global maximum for each bootstrap sample. Thus, the bootstrapped critical value may be much lower than the true critical value (note, however, that Garcia's [53], bootstrapped critical values were considerably higher than his asymptotic critical values). Second, given the unidentified nuisance parameters, the test statistic may not even be asymptotically pivotal. Thus, it is unclear how well the bootstrapped distribution approximates the true finite sample distribution. Despite this, bootstrap tests have often performed better in terms of *size* (the probability of false rejection of the null hypothesis in repeated experiments) than asymptotic tests in the presence of unidentified nuisance parameters. For example, Diebold and Chen [37] consider Monte Carlo analysis of bootstrap and asymptotic tests for structural change with an unknown breakpoint that is a nuisance parameter and find that the bootstrap tests perform well in terms of size and better than the asymptotic tests. Enders, Falk, and Siklos [44] find that bootstrap and asymptotic tests both have size problems for TAR models, although bootstrap LR tests perform better than the asymptotic tests or other bootstrap tests. In terms of testing for nonlinearity with Markov-switching models, Kim, Morley, and Piger [83] bootstrap the distribution of the LR statistic testing linearity for the bounceback model discussed above and reject linearity with a p-value of less than 0.01. The local maximum problem is addressed by conducting a grid search across transition probabilities.

In a recent paper, Carrasco, Hu, and Ploberger [15] develop an information matrix-type test for Markov-switching that is asymptotically optimal and only requires estimation under the null of no Markov-switching (their null allows for other forms of nonlinearity such as ARCH). At this point, there is little known about the finite sample properties of the test. However, Carrasco, Hu, and Ploberger [15] show that it has higher *power* (probability of correct rejection of the null hypothesis in repeated experiments) than Garcia's [53] approach for a basic Markov-switching model with no autoregressive dynamics. Hamilton [60] applies Carrasco, Hu, and Ploberger's [15] method to test for Markov switching in the US unemployment rate (he also provides a very helpful appendix describing how to conduct the test). The null hypothesis is a linear AR(4) model with student t errors. The alternative is an AR(4) with student t errors where the intercept is Markov-switching with three regimes. The test statistic is 26.02, while the 5 percent critical value is 4.01. Thus, linearity can be rejected for the unemployment rate. Meanwhile, the estimated Markov-switching model implies asymmetry in the form of steepness (the unemployment rate rises above its average more quickly than it falls below its average rate).

In contrast to Markov-switching models or threshold models, Beaudry and Koop's [6] ARMA model with the CDR variable provides a very simple test of nonlinearity. In particular, for their preferred specification, Beaudry and Koop [6] find support for nonlinearity with a t-statistic of 3.39 for the CDR variable. Hess and Iwata [68] question the significance of this statistic on the basis of Monte Carlo analysis. However, the data generating process in their Monte Carlo study assumed no drift in the simulated "output" series, meaning that the simulated CDR variable behaves much like a unit root process. By contrast, given drift, the CDR variable can be expected to revert to zero over a fairly short horizon, as it does in the real world (see Fig. 8). Elwood [43] develops an unobserved components model with a threshold process for the transitory component and argues that there is no evidence for asymmetry in the responses to positive and negative shocks. However, his model does not confront the key distinction between large negative shocks versus other shocks that Beaudry and Koop [6] address directly with the inclusion of the CDR variable in their model. A more fundamental issue is whether the CDR variable is merely a proxy for another variable such as the unemployment rate or interest rates and the apparent nonlinearity is simply the result of an omitted variable. However, as discussed in more detail later, the results in Clarida and Taylor [22] and Morley and Piger [113] suggest that Beaudry and Koop's [6] model is capturing a nonlinear dynamic that is fundamentally different than what would be implied by any linear model.

Hess and Iwata [69] provide a more formidable challenge to Beaudry and Koop's [6] model, and, indeed, to many of the regime-switching models discussed earlier, by examining the relative abilities of linear and nonlinear models to reproduce particular features of US real GDP. This alternative form of model evaluation is related to encompassing tests for non-nested models (see [110], on encompassing tests and [9], on the use of encompassing tests to evaluate Markov-switching models). In particular, Hess and Iwata [69] simulate data from a variety of models of output growth, including an AR(1) model, an ARMA(2,2) model, Beaudry and Koop's [6] model, Potter's [121] two-regime TAR model, Pesaran and Potter's [120] "floor and ceiling" model, Hamilton's [59] two-regime Markov-switching model, and a three-regime Markov-switching model with restrictions on the transition matrix as in Boldin [8]. They then consider whether the simulated data for each model can successfully reproduce "business cycle" features in terms of the duration and amplitude of expansions and recessions. Their definition of the business cycle is related to the level of real GDP. However, they label any switch between positive and negative growth, no matter how short-lived, to be a business cycle turning point. For US real GDP, their approach identifies twice as many turning points as reported by the NBER. Under this definition, Hess and Iwata [69] find that the linear AR(1) model is better than the nonlinear models at reproducing the duration and amplitude of "expansions" and "recessions" in US real GDP.

Harding and Pagan [65] and Engel, Haugh, and Pagan [45] confirm Hess and Iwata's [69] findings of little or no "value-added" for nonlinear models over linear models using a business cycle dating procedure that more closely matches NBER dates. The procedure is a quarterly version of an algorithm by Bry and Broschan [12] and identifies recessions as being related to two consecutive quarters of decline in real GDP. In terms of nonlinear models, Engel, Haugh, and Pagan [45] move beyond Hess and Iwata [69] by considering van Dijk and Franses' [149] version of the floor and ceiling model with ARCH errors, Kim, Morley, and Piger's [83] bounceback model, and DeJong, Liesenfeld, and Richard's [34] tension index model. Meanwhile, Clements and Krolzig [25] find that multivariate two-regime Markov-switching models provide little improvement over linear models in capturing business cycle features.

However, beyond the issue of how to define a business cycle, the major question in the literature on reproducing business cycle features is which features to consider. Galvão [52], Kim, Morley, and Piger [83], and Morley and Piger [111] examine the ability of linear and nonlinear models to capture high-growth recoveries that are related to the severity of the preceding recessions, which is the asymmetry emphasized by Friedman [50], Wynne and Balke [150], Sichel [134], and Balke and Wynne [2]. When considering this feature, there is strong support for Kim and Nelson's [79] plucking model and Kim, Morley, and Piger's [83] bounceback model over linear models. Interestingly, the three-regime Markov-switching model does not reproduce this feature. In particular, even though it implies high-growth recoveries, the fixed transition probabilities mean that the strength of the recovery is independent of the severity of the preceding recession. However, the strong support for the plucking model and bounceback model over linear models when considering the relationship between recessions and their recoveries represents a major reversal of the earlier findings for linear models by Hess and Iwata [69] and others.

In terms of directly testing business cycle asymmetries, DeLong and Summers [35] consider a nonparametric test for steepness in real GNP and unemployment rates for eight countries (including the US). In particular, they test for skewness in output growth rates and changes in unemployment rates. With the exception of changes in the US unemployment rate, the measures of economic activity produce no statistically significant evidence of skewness, although the point estimates are generally large and negative for output growth and large and positive for the unemployment rates. Of course, given that the nonparametric test of skewness is unlikely to have much power for the relatively small sample sizes available in macroeconomics, it is hard to treat the non-rejections as particularly decisive. In a more parametric setting, Goodwin [56] considers a likelihood ratio test for sharpness using Hamilton's [59] model. Applying the model and test to real GNP for eight countries (including the US), he is able to reject nonsharpness in every country except Germany. In a more general setting, Clements and Krolzig [24] develop tests of deepness, steepness, and sharpness that are conditional on the number of regimes. For a three-regime model, they are able to reject the null hypotheses of no steepness and no sharpness in US real GDP growth, although the test results are somewhat sensitive to the sample period considered. Meanwhile, Ramsey and Rothman [124] develop a test of time reversibility and find that many measures of economic activity are irreversible and asymmetric, although the nature of the irreversibility does not always provide evidence for nonlinearity.

In addition to classical tests of nonlinear models and the encompassing-style approach discussed above, there are two other approaches to testing nonlinearity that should be briefly mentioned: nonparametric tests and Bayesian model comparison. In terms of nonparametric tests, there is some evidence for nonlinearity in macroeconomic time series. For example, Brock and Sayers [11] apply the nonparametric test for independence (of "prewhitened" residuals using a linear AR model) developed by Brock, Dechert, and Schienkman [10] and are able to reject linearity for the US unemployment rate and industrial production. However, as is always the case with such general tests, it is not clear what alternative is being supported (i. e., is it nonlinearity in the conditional mean or time-variation in the conditional variance?). Also, again, the nonparametric approach is hampered in macroeconomics by relatively small sample sizes. In terms of Bayesian analysis, there is some support for nonlinearity related to business cycle asymmetry using Bayes factors for multivariate models (see [80]). Bayes factors correspond to the posterior odds of one model versus another given equal prior odds. In essence, they compare the relative abilities of two models to predict the data given the stated priors for the model parameters. Obviously, Bayes factors can be sensitive to these priors. However, given diffuse priors, they have a tendency to favor more tightly parametrized models, as some of the prior predictions from the more complicated models can be wildly at odds with the data. Thus, because the findings in favor of nonlinear models correspond to relatively more complicated models, evidence for nonlinearity using Bayes factors is fairly compelling.

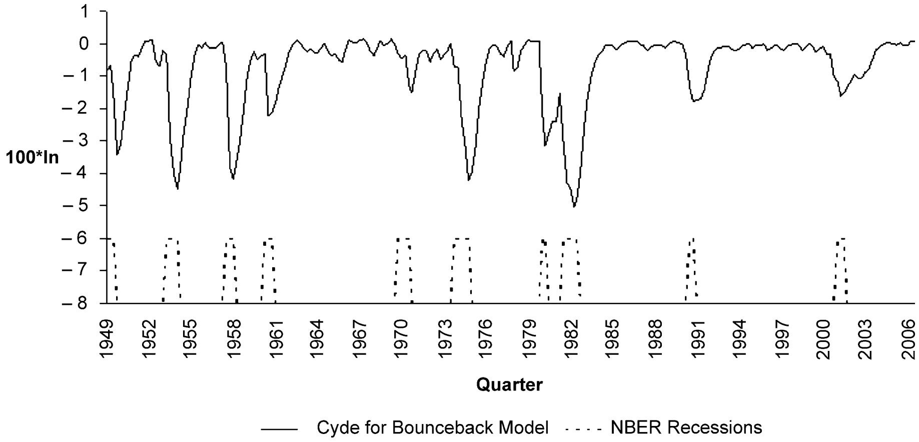
Relevance

Even accepting the presence of nonlinear dynamics related to business cycle asymmetry, there is still a question of economic relevance. Following the literature, the case can be made for relevance in three broad, but related areas: forecasting, macroeconomic policy, and macroeconomic theory.

In terms of forecasting, the nonlinear time series models discussed earlier directly imply different conditional forecasts than linear models. Beaudry and Koop's [6] model provides a simple example with a different implied persistence for large negative shocks than for other shocks. By contrast, linear models imply that the persistence of shocks is invariant to their sign or size. Koop, Pesaran, and Potter [94] develop "generalized impulse response functions" to examine shock dynamics for nonlinear models. Their approach involves simulating artificial time series both in the presence of the shock and in the absence of the shock, holding all else (e. g., other shocks) equal, and comparing the paths of the two simulated time series. This simulation can be done repeatedly for different values of other shocks in order to integrate out their impact on the difference in conditional expectations of the time series implied by presence and absence of a shock. Clarida and Taylor [22] use related simulated forecasts to carry out the Beveridge–Nelson (BN) decomposition (see [7]) for US real GNP using Beaudry and Koop's [6] model. The BN decomposition produces estimates of the permanent and transitory components of a time series based on long-horizon conditional forecasts. Importantly, the estimated cycle (under the "deviations" definition of the business cycle) for Beaudry and Koop's [6] model displays deepness that would be difficult to replicate with any linear forecasting model, even with multivariate information. Thus, there is a direct sense in which Beaudry and Koop's [6] model is not just approximating a linear multivariate model.

In a recent paper, Morley and Piger [112] develop an extension of the BN decomposition that produces optimal (in a "minimum mean squared error" sense) estimates of the cyclical component of an integrated time series when the series can be characterized by a regime-switching process such as for a Markov-switching model with fixed transition probabilities. The approach, which is labeled the "regime-dependent steady-state" (RDSS) decomposition, extracts the trend by constructing a long-horizon forecast conditional on remaining in a particular regime (hence, "regime-dependent"). In Morley and Piger [113], the RDSS decomposition is applied to US real GDP using the "depth" version of Kim, Morley, and Piger's [83] bounceback model given by (8) and (16). Figure 9 presents the estimated cycle for a version of the model with a structural break in σ^2, μ_1, and μ_2 in 1984:Q2 to account for the Great Moderation. The figure also displays an indicator variable for NBER-dated recessions for each quarter from 1949:Q2 to 2006:4. (For visual ease, the indicator variable is -8 in expansions and -6 in recessions).

There are three particularly notable features of the cycle in Fig. 9. First, there is a close correspondence between the big negative movements in it and the NBER-dated periods of recession. Thus, in practice, there is a direct relationship between the level and deviations definitions of the business cycle discussed earlier. Also, this correspondence directly implies that the NBER is identifying a meaningful macroeconomic structure (i. e., it is capturing a phase that is closely related to large movements in the transitory component of economic activity), rather than merely noting negative movements in economic activity. Second, it is fairly evident from Fig. 9 that the cycle displays all three business cycle asymmetries in the form of deepness, steepness, and sharpness. Third, the unconditional mean of the cycle is negative. As discussed in Morley and Piger [113],



Macroeconomics, Non-linear Time Series in, Figure 9
"Bounceback" cycle and NBER recessions (Source: Author's calculations based on Morley and Piger [113], and NBER website)

this finding stands in contrast to cyclical estimates for all linear models, whether univariate or multivariate.

The negative mean of the cycle in US real GDP has strong implications for the potential benefits of macroeconomic stabilization policy. Lucas [102,103] famously argued that the elimination of all business cycle fluctuations would produce a benefit equivalent to less than one-tenth of one percent of lifetime consumption. One reason for this extraordinarily low estimate is that his calculation assumes business cycle fluctuations are symmetric. However, as discussed in DeLong and Summers [36], Cohen [28], Barlevy [5], and Yellen and Akerlof [151], a non-zero mean cyclical component of economic activity directly implies that stabilization policies, if effective, could raise the average level of output and lower the average level of the unemployment rate. In this setting, the potential benefits of stabilization policy are much larger than calculated by Lucas [102,103]. (In deference to Milton Friedman and his plucking model, it is worth mentioning that the optimal "stabilization" policy might be a passive rule that prevents policymakers from generating recessionary shocks in the first place. Regardless, the point is that, given a negative mean for the cycle in real GDP, the costs of business cycles are high and can be affected by policy).

A related issue is asymmetry in terms of the effects of macroeconomic policy on economic activity. For example, DeLong and Summers [36] and Cover [31] find that negative monetary policy shocks have a larger effect on output than positive shocks of the same size (the so-called "pushing on a string" hypothesis). This form of asymmetry represents a third type of nonlinearity in macroeconomics beyond structural change and business cycle asymmetry, although it is clearly related to business cycle asymmetry. Indeed, Garcia and Schaller [54] and Lo and Piger [99] consider Markov-switching models and find that asymmetry in the effects of monetary policy shocks is more closely related to whether the economy is in an expansion or a recession, rather than whether the shock was positive or negative. In particular, positive shocks can have large effects on output, but only in recessions. There is an obvious link between this result, which is suggestive of a convex short-run aggregate supply curve rather than the "pushing on a string" hypothesis, and the business cycle displayed in Fig. 9, which is also highly suggestive of a convex short-run aggregate supply curve.

In addition to the implications for more traditional theoretical notions in macroeconomics such as the shape of the short-run aggregate supply curve, the findings for business cycle asymmetry are important for modern macroeconomic theory because dynamic stochastic general equilibrium (DSGE) models are often evaluated and compared based on their ability to generate internal propagation that matches what would be implied by linear AR and VAR models of US real GDP (see, for example [126]). These linear models imply a time-invariant propagation structure for shocks, while the business cycle presented in Fig. 9 suggests that theory-based models should instead be evaluated on their ability to generate levels of propagation that vary over business cycle regimes, at least if they are claimed to be "business cycle" models.

Future Directions

There are several interesting avenues for future research in nonlinear time series in macroeconomics. However, two follow directly from the findings on nonlinearities summarized in this survey. First, in terms of structural change, it would be useful to determine whether the process of change is gradual or abrupt and the extent to which it is predictable. Second, in terms of business cycle asymmetries, it would be useful to pin down the extent to which they reflect nonlinearities in conditional mean dynamics, conditional variance dynamics, and/or the contemporaneous relationship between macroeconomic variables.

The issue of whether structural change is gradual or abrupt is only meaningful when structural change is thought of as a form of nonlinearity in a time series model. In particular, formal classical tests of structural change based on asymptotic theory make no distinction between whether there are many small change or a few large changes. All that matters is the cumulative magnitude of changes over the long horizon (see [42], on this point). Of course, a time-varying parameter model and a regime-switching model with permanent changes in regimes can fit the data in very different ways in finite samples. Thus, it is possible to use finite-sample model comparison (e. g., Bayes factors) to discriminate between these two behaviors. It is even possible to use a particle filter to estimate a nonlinear state-space model that nests large, infrequent changes and small, frequent changes (see [90]). In terms of predicting structural change, Koop and Potter [93] develop a flexible model that allows the number of structural breaks in a given sample and the duration of structural regimes to be stochastic processes and discuss estimation of the model via Bayesian methods.

The issue of the relative importance of different types of recurring nonlinearities is brought up by the findings in Sims and Zha [136], discussed earlier, that there are no changes in the conditional mean dynamics, but only changes in the conditional variance of shocks for a structural VAR model of the US economy. Likewise, in their multivariate three-regime TAR model, Koop and Potter [92] consider a VAR structure, and find that a linear VAR structure with heteroskedastic errors is preferred over a "vector floor and ceiling" structure for the conditional mean dynamics. The question is how to reconcile these results with the large body of evidence supporting nonlinearity in conditional mean dynamics discussed at length in this survey. A short answer is that VAR models are highly parametrized in terms of the conditional mean. Thus, it may be hard to identify regime shifts or nonlinear forms of time-variation in conditional means using a VAR model, even if they are present. On the other hand, even for their nonlinear model, Koop and Potter [92] find stronger evidence for nonlinearity in the contemporaneous relationship between variables than in the conditional mean dynamics. Meanwhile, in terms of multivariate analysis, consideration of more parsimonious factor models has typically increased the support for nonlinear models over linear models (e. g. [80]). Thus, a full comparison of different types of nonlinearity in the context of a parsimonious nonlinear model would be useful.

Another important avenue for future research in macroeconomics is an increased integration of the findings in nonlinear time series into macroeconomic theory. In terms of structural change, there has been considerable progress in recent years. In particular, some of the papers on changes in policy regimes discussed earlier (e. g. [123, 136]) can be classified as"theory-oriented" given their consideration of structural VAR models. Another nonlinear time series paper on changing policy regimes with a structural model is Owyang and Ramey [118], which considers the interaction between regime switching in the Phillips curve and the policy rule. Meanwhile, Fernández-Villaverde and Rubio-Ramírez [47] and King [89] directly incorporate structural change (of the gradual form) in theory-based DSGE models, which they proceed to estimate with the aid of particle filters. In terms of Bayesian analysis of the sources of the Great Moderation, Chauvet and Potter [20] and Kim, Nelson, and Piger [82] consider disaggregated data (in a joint model and separately, respectively) and find that the decline in volatility of economic activity is a broadly-based phenomenon, rather than corresponding to particular sectors, while Kim, Morley, and Piger [86] employ structural VAR models and find that the decline in volatility cannot be explained by changes in aggregate demand shocks, monetary policy shocks, or the response of the private sector or policymakers to shocks.

In terms of the integration of business cycle asymmetries into macroeconomic theory, there has been less progress in recent years, perhaps due the obviously greater difficulty in modeling endogenous regime switching than in simply assuming exogenous structural change. However, the theoretical literature contains some work on asymmetries. In particular, mechanisms for regime switching in the aggregate data that have been considered in the past include spillovers and strategic complementarities [41], animal spirits [70], a history-dependent selection criterion in an economy with multiple Nash equilibria corresponding to different levels of productivity [30], and intertemporal increasing returns [1]. However, Potter [122] notes that, while these mechanisms can generate regime switching in the aggregate data, they cannot explain asymmetry in the form of high-growth recoveries following large negative shocks. He proposes a model with Bayesian learning and an information externality (see [16]) that can generate such dynamics. Meanwhile, in terms of business cycle asymmetry more generally, obvious mechanisms are investment irreversibilities [55] and capacity constraints [64]. More promisingly for future developments in macroeconomic theory, there is a growing empirical literature on the sources of business cycle asymmetries. For example, Korenok, Mizrach, and Radchenko [95] use disaggregated data and find that asymmetries are more pronounced in durable goods manufacturing sectors than nondurable goods manufacturing sectors (also see [129]) and appear to be related to variation across sectors in credit conditions and reliance on raw material inventories, while they do not appear to be related to oil price shocks [33] or adjustment costs [3].

Bibliography

Primary Literature

1. Acemoglu D, Scott A (1997) Asymmetric business cycles: Theory and time-series evidence. J Monet Econ 40:501–533
2. Balke NS, Wynne MA (1996) Are deep recessions followed by strong recoveries? Results for the G-7 countries. Appl Econ 28:889–897
3. Ball L, Mankiw NG (1995) Relative price changes as aggregate supply shocks. Q J Econ 110:161–193
4. Bansal R, Yaron A (2004) Risks for the long run: A potential resolution of asset pricing puzzles. J Financ 59:1481–1509
5. Barlevy G (2005) The cost of business cycles and the benefits of stabilization. Econ Perspect 29:32–49
6. Beaudry P, Koop G (1993) Do recessions permanently change output? J Monet Econ 31:149–163
7. Beveridge S, Nelson CR (1981) A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. J Monet Econ 7:151–174
8. Boldin MD (1996) A check on the robustness of Hamilton's Markov switching model approach to the economic analysis of the business cycle. Stud Nonlinear Dyn Econom 1:35–46

9. Breunig R, Najarian S, Pagan A (2003) Specification testing of Markov-switching models. Oxf Bull Econ Stat 65:703–725
10. Brock WA, Dechert WD, Scheinkman JA (1996) A test of independence based on the correlation dimension. Econom Rev 15:197–235
11. Brock WA, Sayers C (1988) Is the business cycle characterized by deterministic chaos? J Monet Econ 22:71–90
12. Bry G, Boschan C (1971) Cyclical analysis of time series: Selected procedures and computer programs. NBER, New York
13. Burns AF, Mitchell WA (1946) Measuring Business Cycles. NBER, New York
14. Camacho M (2005) Markov-switching stochastic trends and economic fluctuations. J Econ Dyn Control 29:135–158
15. Carrasco M, Hu L, Ploberger W (2007) Optimal test for Markov switching. Working Paper
16. Chalkley M, Lee IH (1998) Asymmetric business cycles. Rev Econ Dyn 1:623–645
17. Chan KS (1991) Percentage points of likelihood ratio tests for threshold autoregression. J Royal Stat Soc Ser B 53:691–696
18. Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Tim Ser Analysis 7:179–190
19. Chauvet M (1998) An econometric characterization of business cycle dynamics with factor structure and regime switches. Int Econ Rev 39:969–996
20. Chauvet M, Potter S (2001) Recent changes in the US business cycle. Manch Sch 69:481–508
21. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
22. Clarida RH, Taylor MP (2003) Nonlinear permanent-temporary decompositions in macroeconomics and finance. Econ J 113:C125–C139
23. Clements MP, Krolzig HM (1998) A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. Econ J 1:C47–C75
24. Clements MP, Krolzig HM (2003). Business cycle asymmetries: Characterization and testing based on Markov-switching autoregressions. J Bus Econ Stat 21:196–211
25. Clements MP, Krolzig HM (2004) Can regime-switching models reproduce the business cycle features of US aggregate consumption, investment and output? Int J Financ Econ 9:1–14
26. Cogley T, Sargent TJ (2001) Evolving post-World War II US inflation dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 331–373
27. Cogley T, Sargent TJ (2005) Drift and volatilities: Monetary policies and outcomes in the post WW II US. Rev Econ Dyn 8:262–302
28. Cohen D (2000) A quantitative defense of stabilization policy. Federal Reserve Board Finance and Economics Discussion Series. Paper 2000-34
29. Cooley TF, Prescott EC (1976) Estimation in the presence of stochastic parameter variation. Econometrica 44:167–184
30. Cooper R (1994) Equilibrium selection in imperfectly competitive economies with multiple equilibria. Econ J 104:1106–1122
31. Cover JP (1992) Asymmetric effects of positive and negative money-supply shocks. Q J Econ 107:1261–1282
32. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254
33. Davis SJ, Haltiwanger J (2001) Sectoral job creation and destruction responses to oil price changes. J Monet Econ 48:468–512
34. DeJong DN, Liesenfeld R, Richard JF (2005) A nonlinear forecasting model of GDP growth. Rev Econ Stat 87:697–708
35. DeLong JB, Summers LH (1986) Are business cycles symmetrical? In: Gordon RJ (ed) The American Business Cycle. University of Chicago Press, Chicago, pp 166–179
36. DeLong B, Summers L (1988) How does macroeconomic policy affect output? Brook Papers Econ Activity 2:433–480
37. Diebold FX, Chen C (1996) Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. J Econ 70:221–241
38. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. Rev Econ Stat 78:67–77
39. Diebold FX, Lee JH, Weinbach G (1994) Regime switching with time-varying transition probabilities. In: Hargreaves C (ed) Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford, pp 283–302
40. Durland JM, McCurdy TH (1994) Duration-dependent transitions in a Markov model of US GNP growth. J Bus Econ Stat 12:279–288
41. Durlauf SN (1991) Multiple equilibria and persistence in aggregate fluctuations. Am Econ Rev Pap Proc 81:70–74
42. Elliott G, Müller U (2006) Efficient tests for general persistent time variation in regression coefficients. Rev Econ Stud 73:907–940
43. Elwood SK (1998) Is the persistence of shocks to output asymmetric? J Monet Econ 41:411–426
44. Enders W, Falk BL, Siklos P (2007) A threshold model of real US GDP and the problem of constructing confidence intervals in TAR models. Stud Nonlinear Dyn Econ 11(3):4
45. Engel J, Haugh D, Pagan A (2005) Some methods for assessing the need for non-linear models in business cycles. Int J Forecast 21:651–662
46. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
47. Fernández-Villaverde J, Rubio-Ramírez JF (2007) Estimating macroeconomic models: A likelihood approach. Rev Econ Stud 54:1059–1087
48. Filardo AJ (1994) Business-cycle phases and their transitional dynamics. J Bus Econ Stat 12:299–308
49. French MW, Sichel DE (1993) Cyclical patterns in the variance of economic activity. J Bus Econ Stat 11:113–119
50. Friedman M (1964) Monetary Studies of the National Bureau, the National Bureau Enters Its 45th Year. 44th Annual Report. NBER, New York, pp 7–25; Reprinted in: Friedman M (1969) The Optimum Quantity of Money and Other Essays. Aldine, Chicago, pp 261–284
51. Friedman M (1993) The "plucking model" of business fluctuations revisited. Econ Inq 31:171–177
52. Galvão AB (2002) Can non-linear time series models generate US business cycle asymmetric shape? Econ Lett 77:187–194
53. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int Econ Rev 39:763–788
54. Garcia R, Schaller H (2002) Are the effects of interest rate changes asymmetric? Econ Inq 40:102–119

55. Gilchrist S, Williams JC (2000) Putty-clay and investment: A business cycle analysis. J Political Econ 108:928–960
56. Goodwin TH (1993) Business-cycle analysis with a Markov-switching model. J Bus Econ Stat 11:331–339
57. Granger CWJ, Andersen AP (1978) An Introduction to Bilinear Time Series Models. Vandenhoek and Ruprecht, Göttingen
58. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford
59. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
60. Hamilton JD (2005) What's real about the business cycle? Fed Reserve Bank St. Louis Rev 87:435–452
61. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. J Appl Econ 7:S61–S82
62. Hansen BE (1996) Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64:413–430
63. Hansen BE (1997) Inference in TAR models. Stud Nonlinear Dyn Econom 2:1–14
64. Hansen GD, Prescott EC (2005) Capacity constraints, asymmetries, and the business cycle. Rev Econ Dyn 8:850–865
65. Harding D, Pagan AR (2002) Dissecting the cycle: A methodological investigation. J Monet Econ 49:365–381
66. Harding D, Pagan AR (2003) A Comparison of Two Business Cycle Dating Methods. J Econ Dyn Control 27:1681–1690
67. Harding D, Pagan AR (2005) A suggested framework for classifying the modes of cycle research. J Appl Econom 20:151–159
68. Hess GD, Iwata S (1997) Asymmetric persistence in GDP? A deeper look at depth. J Monet Econ 40:535–554
69. Hess GD, Iwata S (1997) Measuring and comparing business-cycle features. J Bus Econ Stat 15:432–444
70. Howitt P, McAfee RP (1992) Animal spirits. Am Econ Rev 82:493–507
71. Hristova D (2005) Maximum likelihood estimation of a unit root bilinear model with an application to prices. Stud Nonlinear Dyn Econom 9(1):4
72. Keynes JM (1936) The General Theory of Employment, Interest, and Money. Macmillan, London
73. Kiefer NM (1978) Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica 46:413–430
74. Kim CJ (1994) Dynamic linear models with Markov switching. J Econom 60:1–22
75. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. Empir Econ 27:163–183
76. Kim CJ, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. Rev Econ Stat 80:188–201
77. Kim CJ, Nelson CR (1999) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
78. Kim CJ, Nelson CR (1999) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econ Stat 81:608–616
79. Kim CJ, Nelson CR (1999) Friedman's plucking model of business fluctuations: Tests and estimates of permanent and transitory components. J Money Credit Bank 31:317–34
80. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. Int Econ Rev 42:989–1013
81. Kim CJ, Piger JM (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1181–1211
82. Kim CJ, Nelson CR, Piger J (2004) The less-volatile US economy: A Bayesian investigation of timing, breadth, and potential explanations. J Bus Econ Stat 22:80–93
83. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the permanent effects of recessions. J Appl Econom 20:291–309
84. Kim CJ, Piger J, Startz R (2007) The dynamic relationship between permanent and transitory components of US business cycles. J Money Credit Bank 39:187–204
85. Kim CJ, Piger J, Startz R (2008) Estimation of Markov regime-switching regression models with endogenous switching. J Econom 143:263–273
86. Kim CJ, Morley J, Piger J (2008) Bayesian Counterfactual Analysis of the Sources of the Great Moderation. J Appl Econom 23:173–191
87. Kim M-J, Yoo J-S (1995) New index of coincident indicators: A multivariate Markov switching factor model approach. J Monet Econ 36:607–630
88. Kim S, Shephard N, Chib S (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models. Rev Econ Stud 65:361–393
89. King TB (2006) Dynamic equilibrium models with time-varying structural parameters. Working Paper
90. King TB, Morley J (2007) Maximum likelihood estimation of nonlinear, non-Gaussian state-space models using a multi-stage adaptive particle filter. Working Paper
91. Koop G, Potter S (2003) Bayesian analysis of endogenous delay threshold models. J Bus Econ Stat 21:93–103
92. Koop G, Potter S (2006) The vector floor and ceiling model. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 97–131
93. Koop G, Potter S (2007) Estimation and forecasting in models with multiple breaks. Rev Econ Stud 74:763–789
94. Koop G, Pesaran MH, Potter S (1996) Impulse response analysis in nonlinear multivariate models. J Econometrics 74:119–148
95. Korenok O, Mizrach B, Radchenko S (2009) A note on demand and supply factors in manufacturing output asymmetries. Macroecon Dyn (forthcoming)
96. Lam PS (1990) The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. J Monet Econ 26:409–432
97. Leamer EE, Potter SM (2004) A nonlinear model of the business cycle. Working Paper
98. Leyton AP, Smith D (2000) A further note of the three phases of the US business cycle. Appl Econ 32:1133–1143
99. Lo MC, Piger J (2005) Is the response of output to monetary policy asymmetric? Evidence from a regime-switching coefficients model. J Money Credit Bank 37:865–887
100. Lucas RE (1972) Econometric testing of the natural rate hypothesis. In: Eckstein O (ed) Econometrics of Price Determination. US Federal Reserve Board, Washington DC, pp 50–59
101. Lucas RE (1976) Econometric policy evaluation: A critique. In: Brunner K, Meltzer A (eds) The Phillips Curve and Labor Markets, vol 1. Carnegie-Rochester Ser Public Policy, pp 19–46

102. Lucas RE (1987) Models of Business Cycles. Basil Blackwell, Oxford
103. Lucas RE (2003) Macroeconomic Priorities. Am Econ Rev 93:1–14
104. Ma J (2007) Consumption persistence and the equity premium puzzle: New evidence based on improved inference. Working paper
105. MacKinnon J (2002) Bootstrap inference in econometrics. Can J Econ 35:615–645
106. MacKinnon J (2006) Bootstrap methods in econometrics. Econ Rec 82:S2–S18
107. McConnell MM, Quiros GP (2000) Output fluctuations in the United States: What has changed since the early 1980s? Am Econ Rev 90:1464–1476
108. McQueen G, Thorley SR (1993) Asymmetric business cycle turning points. J Monet Econ 31:341–362
109. Mitchell WA (1927) Business Cycles: The Problem and Its Setting. NBER, New York
110. Mizon GE, Richard JF (1986) The encompassing principle and its application to non-nested hypotheses. Econometrica 54:657–678
111. Morley J, Piger J (2006) The Importance of Nonlinearity in Reproducing Business Cycle Features. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 75–95
112. Morley J, Piger J (2008) Trend/cycle decomposition of regime-switching processes. J Econom (forthcoming)
113. Morley J, Piger J (2008) The asymmetric business cycle. Working Paper
114. Morley JC, Nelson CR, Zivot E (2003) Why are the Beveridge-Nelson and unobserved-components decompositions of GDP so different? Rev Econ Stat 85:235–243
115. Neftçi SH (1984) Are economic time series asymmetric over the business cycle? J Political Econ 92:307–328
116. Niemira MP, Klein PA (1994) Forecasting Financial and Economic Cycles. Wiley, New York
117. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. J Appl Econom 15:27–44
118. Owyang MT, Ramey G (2004) Regime switching and monetary policy measurement. J Monet Econ 51:1577–1198
119. Peel D, Davidson J (1998) A non-linear error correction mechanism based on the bilinear model. Econ Lett 58:165–170
120. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
121. Potter SM (1995) A nonlinear approach to US GNP. J Appl Econ 10:109–125
122. Potter SM (2000) A nonlinear model of the business cycle. Stud Nonlinear Dyn Econom 4:85–93
123. Primiceri GE (2005) Time varying structural vector autogressions and monetary policy. Rev Econ Stud 72:821–852
124. Ramsey JB, Rothman P (1996) Time irreversibility and business cycle asymmetry. J Money Credit Bank 28:1–21
125. Ravn MO, Sola M (1995) Stylized facts and regime changes: Are prices procyclical? J Monet Econ 36:497–526
126. Rotemberg JJ, Woodford M (1996) Real-business-cycle Models and the forecastable movements in output, hours, and consumption. Am Econ Rev 86:71–89
127. Rothman P (1991) Further Evidence on the Asymmetric Behavior of Unemployment Rates Over the Business Cycle. J Macroeconom 13:291–298
128. Rothman P (1998) Forecasting asymmetric unemployment rates. Rev Econ Stat 80:164–168
129. Rothman P (2008) Reconsideration of Markov chain evidence on unemployment rate asymmetry. Stud Nonlinear Dyn Econo 12(3):6
130. Rothman P, van Dijk D, Franses PH (2001) A multivariate STAR analysis of the relationship between money and output. Macroeconom Dyn 5:506–532
131. Schumpeter J (1942) Capitalism, socialism, and democracy. Harper, New York
132. Sensier M, van Dijk D (2004) Testing for volatility changes in US macroeconomic time series. Rev Econ Stat 86:833–839
133. Sichel DE (1993) Business cycle asymmetry: A deeper look. Econ Inq 31:224–236
134. Sichel DE (1994) Inventories and the three phases of the business cycle. J Bus Econ Stat 12:269–277
135. Sims CA (2001) Comment on Sargent and Cogley's: Evolving Post-World War II US Inflation Dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 373–379
136. Sims CA, Zha T (2006) Were there regime switches in US monetary policy? Am Econ Rev 96:54–81
137. Sinclair TM (2008) Asymmetry in the business cycle: Friedman's plucking model with correlated innovations. Working Paper
138. Stock JH, Watson MW (2002) Has the business cycle changed and why? In: Gertler M, Rogoff K (eds) NBER Macroeconomics Annual 2002. MIT Press, Cambridge, pp 159–218
139. Subba Rao T, Gabr MM (1984) An Introduction to Bispectral Analysis and Bilinear Time Series Models. Lecture Notes in Statistics, vol 24. Springer, New York
140. Teräsvirta T (1994) Specification, estimation, and evaluation of smooth transition autoregressive models. J Am Stat Assoc 89:208–218
141. Teräsvirta T (1995) Modeling nonlinearity in US Gross National Product 1889–1987. Empir Econ 20:577–598
142. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp 507–552
143. Teräsvirta T, Anderson HM (1992) Characterizing nonlinearities in business cycles using smooth transition autoregressive models. J Appl Econom 7:S119–S136
144. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time- series analysis. J Forecast 13:109–131
145. Tong H (1978) On a threshold model. In: Chen CH (ed) Pattern Recognition and Signal Processing. Sijhoff and Noordhoff, Amsterdam, pp 575–586
146. Tsay RS (1989) Testing and modeling threshold autoregressive processes. J Am Stat Assoc 84:231–240
147. Tsay RS (1998) Testing and modeling multivariate threshold processes. J Am Stat Assoc 93:1188–1202
148. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroeconom Dyn 3:311–340
149. van Dijk D, Franses PH (2003) Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. Oxf Bull Econ Stat 65:727–744

150. Wynne MA, Balke NS (1992) Are deep recessions followed by strong recoveries? Econ Lett 39:183–189
151. Yellen JL, Akerlof GA (2006) Stabilization policy: A reconsideration. Econ Inq, pp 44:1–22

Books and Reviews

Davidson R, MacKinnon JG (2004) Econometric Theory and Methods. Oxford University Press, Oxford

Diebold FX (1998) The past, present, and future of macroeconomic forecasting. J Econ Perspectives 12:175–192

Engle R (2001) GARCH 101: The use of ARCH/GARCH models in applied econometrics. J Econc Perspectives 15:157–168

Franses PH (1998) Time Series Models for Business and Economic Forecasting. Cambridge University Press, Cambridge

Hamilton JD (1994) State-space models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 041–3080

Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton

Koop G (2003) Bayesian Econometrics. Wiley, Chichester

Teräsvirta T, Tjøstheim D, Granger CWJ (1994) Aspects of modeling nonlinear time series. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 2919–2957

Tsay RS (2005) Analysis of Financial Time Series. Wiley, Hoboken

Market Games and Clubs

Myrna Wooders
Department of Economics, Vanderbilt University,
Nashville, USA

Article Outline

Glossary

Game A (cooperative) game (in characteristic form) is defined simply as a finite set of players and a function or correspondence ascribing a worth (a non-negative real number, interpreted as an idealized money) to each nonempty subset of players, called a group or coalition.

Payoff vector A payoff vector is a vector listing a payoff (an amount of utility or money) for each player in the game.

Core The core of a game is the set (possibly empty) of feasible outcomes – divisions of the worths arising from coalition formation among the players of the game – that cannot be improved upon by any coalition of players. core

Totally balanced game A game is totally balanced if the game and every subgame of the game (a game with player set taken as some subset of players of the initially given game) has a nonempty core.

Market A market is defined as a private goods economy in which all participants have utility functions that are linear in (at least) one commodity (money).

Shapley value The Shapley value of a game is feasible outcome of a game in which all players are assigned their expected marginal contribution to a coalition when all orders of coalition formation are equally likely.

Pregame A pair, consisting of a set of player types (attributes or characteristics) and a function mapping finite lists of characteristics (repetitions allowed) into the real numbers. In interpretation, the pregame function ascribes a worth to every possible finite group of players, where the worth of a group depends on the numbers of players with each characteristic in the group. A pregame is used to generate games with arbitrary numbers of players.

Small group effectiveness A pregame satisfies small group effectiveness if almost all gains to collective activities can be realized by cooperation only within arbitrarily small groups (coalitions) of players.

Per capita boundedness A pregame satisfies per capita boundedness if the supremum of the average worth of any possible group of players (the per capita payoff) is finite.

Asymptotic negligibility A pregame satisfies asymptotic negligibility if vanishingly small groups can have only negligible effects on per capita payoffs.

Market games A market game is a game derived from a market. Given a market and a group of agents we can determine the total utility (measured in money) that the group can achieve using only the endowments belonging to the group members, thus determining a game.

Club A club is a group of agents or players that forms for the purpose of carrying out come activity, such as providing a local public good.

An economy We use the term ‘economy’ to describe any economic setting, including economies with clubs, where the worth of club members may depend on the characteristics of members of the club, economies with pure public goods, local public goods (public goods subject to crowding and/or congestion), economies with production where what can be produced and the costs of production may depend on the characteristics of the individuals involved in production, and so on. A *large economy* has many participants.

Price taking equilibrium A price taking equilibrium for a market is a set of prices, one for each commodity, and an allocation of commodities to agents so that each agent can afford his part of the allocation, given the value of his endowment.

Definition of the Subject

The equivalence of markets and games concerns the relationship between two sorts of structures that appear fundamentally different – markets and games. Shapley and Shubik [60] demonstrates that: (1) games derived from markets with concave utility functions generate totally balanced games where the players in the game are the participants in the economy and (2) every totally balanced game generates a market with concave utility functions. A particular form of such a market is one where the commodities are the participants themselves, a labor market for example.

But markets are very special structures, more so when it is required that utility functions be concave. Participants may also get utility from belonging to groups, such as marriages, or clubs, or productive coalitions. It may be that participants in an economy even derive utility (or disutility) from engaging in processes that lead to the eventual exchange of commodities. The question is when are such economic structures equivalent to markets with concave utility functions.

This paper summarizes research showing that a broad class of large economies generate balanced market games. The economies include, for example, economies with clubs where individuals may have memberships in multiple clubs, with indivisible commodities, with nonconvexities and with non-monotonicities. The main assumption are: (1) that an option open to any group of players is to break into smaller groups and realize the sum of the worths of these groups, that is, essential superadditivity is satisfied and: (2) relatively small groups of participants can realize almost all gains to coalition formation.

The equivalence of games with many players and markets with many participants indicates that relationships obtained for markets with concave utility functions and many participants will also hold for diverse social and economic situations with many players. These relationships include: (a) equivalence of the core and the set of competitive outcomes; (b) the Shapley value is contained in the core or approximate cores; (c) the equal treatment property holds – that is, both market equilibrium and the core treat similar players similarly. These results can be applied to diverse economic models to obtain the equivalence of cooperative outcomes and competitive, price taking outcomes in economies with many participants and indicate that such results hold in yet more generality.

Introduction

One of the subjects that has long intrigued economists and game theorists is the relationship between games, both cooperative and noncooperative, and economies. Seminal works making such relationships include Shubik [67], Debreu and Scarf [22], Aumann [4], Shapley and Shubik [60,62] and Aumann and Shapley [7], all connecting outcomes of price-taking behavior in large economies with cores of games. See also Shapley and Shubik [63] and an ongoing stream of papers connecting strategic behavior to market behavior. Our primary concern here, however, is not with the equivalence of outcomes of solution concepts for economies, as is Debreu and Scarf [22] or Aumann [6] for example, but rather with equivalences of the *structures* of markets and games. Solution concepts play some role, however, in establishing these equivalences and in understanding the meaning of the equivalence of markets and games.

In this entry, following Shapley and Shubik [60], we focus on markets in which utility functions of participants are quasi-linear, that is, the utility function u of a participant can be written as $u(x,\xi) = \widehat{u}(x) + \xi$ where $x \in \mathbb{R}^L_+$ is a commodity bundle, $\xi \in \mathbb{R}$ is interpreted as money and $\widehat{u}$ is a continuous function. Each participant in an economy has an endowment of commodities and, without any substantive loss of generality, it is assumed that no money is initially endowed. The price of money is assumed equal to one. A price taking equilibrium for a market then consists of a price vector $p \in \mathbb{R}^L$ for the commodities.and an assignment of commodities to participants such that: the total amounts of commodities assigned to participants equals the total amount of commodities with which participants are endowed and; given prices, each participant can afford his assignment of commodities and no participant, subject to his budget constraint, can afford a preferred commodity bundle.

We also treat games with side payments, alternatively called games with transferable utility or, in brief, TU games. Such a game consists of a finite set N of players and a worth function that assigns to each group of players $S \subset N$ a real number $v(S) \in \mathbb{R}_+$, called the worth of the group. In interpretation, $v(S)$ is the total payoff that a group of players can realize by cooperation. A central game-theoretic concept for the study of games is the core. The core consists of those divisions of the maximal total worth achievable by cooperation among the players in N so that each group of players is assigned at least its worth. A game is balanced if it has a nonempty core and totally balanced if all subgames of the game have nonempty cores. A subgame of a game is simply a group of players $S \subset N$ and the worth function restricted to that group and the smaller groups that it contains.

Given a market any feasible assignment of commodities to the economic participants generates a total worth of each group of participants. The worth of a group of participants (viewed as players of a game) is the maximal total utility achievable by the members of the group by allocating the commodities they own among themselves. In this way a market generates a game – a set of players (the participants in the economy) and a worth for each group of players.

Shapley and Shubik [60] demonstrate that any market where all participants have concave, monotonic increasing utility functions generates a totally balanced game and that any totally balanced game generates a market, thus establishing an equivalence between a class of markets and totally balanced cooperative games. A particular sort of market is canonical; one where each participant in the market is endowed with one unit of a commodity, his "type". Intuitively, one might think of the market as one where each participant owns one unit of himself or of his labor.

In the last twenty years or so there has been substantial interest in broader classes of economies, including those with indivisibilities, nonmonotonicities, local public goods or clubs, where the worth of a group depends not only on the private goods endowed to members of the group but also on the characteristics of the group members. For example, the success of the marriage of a man and a woman depends on their characteristics and on whether their characteristics are complementary. Similarly, the output of a machine and a worker using the machine depends on the quality and capabilities of the machine and how well the abilities of the worker fit with the characteristics of the machine – a concert pianist fits well with an high quality piano but perhaps not so well with a sewing machine. Or how well a research team functions depends not only on the members of the team but also on how well they interact. For simplicity, we shall refer to these economies as club economies. Such economies can be modeled as cooperative games.

In this entry we discuss and summarize literature showing that economies with many participants are approximated by markets where all participants have the same concave utility function and for which the core of the game is equivalent to the set of price-taking economic equilibrium payoffs. The research presented is primarily from Shubik and Wooders [65], Wooders [92] and earlier papers due to this author. For the most recent results in this line of research we refer the reader to Wooders [93,94,95]. We also discuss other related works throughout the course of the entry. The models and results are set in a broader context in the conclusions.

The importance of the equivalence of markets and games with many players relates to the hypothesis of perfect competition, that large numbers of participants leads to price-taking behavior, or behavior "as if" participants took prices as given. Von Neumann and Morgenstern perceived that even though individuals are unable to influence market prices and cannot benefit from strategic behavior in large markets, large "coalitions" might form. Von Neumann and Morgenstern write:

> It is neither certain nor probable that a mere increase in the number of participants might lead *in fine* to the conditions of free competition. The classical definitions of free competition all involve further postulates besides this number. E.g., it is clear that if certain great groups of individuals will – for any reason whatsoever– act together, then the great number of participants may not become effective; the decisive exchanges may take place directly between large "coalitions", few in number and not between individuals, many in number acting independently. … Any satisfactory theory … will have to explain when such big coalitions will or will not be formed –i. e., when the large numbers of participants will become effective and lead to more or less free competition.

The assumption that small groups of individuals cannot affect market aggregates, virtually taken for granted by von Neumann and Morgenstern, lies behind the answer to the question they pose. The results presented in this entry suggest that the great number of participants will become effective and lead to more or less free competition when *small* groups of participants cannot significantly affect market outcomes. Since all or almost all gains to collective activities can be captured by relatively small groups, large groups gain no market power from size; in other words, large groups are inessential. That large groups are inessential is equivalent to small group effectiveness [89]. A remarkable feature of the results discussed in this essay is they are independent of any particular economic structure.

Transferable Utility Games; Some Standard Definitions

Let (N, ν) be a pair consisting of a finite set N, called a *player set*, and a function ν, called a *worth function*, from subsets of N to the real numbers $\mathbb{R}$ with $\nu(\phi) = 0$. The pair (N, ν) is a *TU game* (also called a game with side payments). Nonempty subsets S of N are called *groups* (of players) and the number of members of the group S is given by $|S|$. Following is a simple example.

Example 1 A glove game: Suppose that we can partition a player set N into two groups, say N_1 and N_2. In interpretation, a member of N_1 is endowed with a right-hand (RH) glove and a member of N_2 is endowed with a left-hand

(LH) glove. The worth of a pair of gloves is \$1, and thus the worth of a group of players consisting of player $i \in N_1$ and player $j \in N_2$ is \$1. The worth of a single glove and hence of a one-player group is \$0. The worth of a group $S \subset N$ is given by $v(S) = \min\{|S \cap N_1|, |S \cap N_2|\}$. The pair (N, v) is a game.

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. We regard vectors in finite dimensional Euclidean space $\mathbb{R}^T$ as functions from T to $\mathbb{R}$, and write $\overline{u}_i$ for the ith component of $\overline{u}$, etc. If $S \subset T$ and $\overline{u} \in \mathbb{R}^T$, we shall write $\overline{u}_S := (\overline{u}_i : i \in S)$ for the restriction of $\overline{u}$ to S. We write 1_S for the element of $\mathbb{R}^S$ all of whose coordinates are 1 (or simply 1 if no confusion can arise.) A payoff vector $\overline{u}$ is *feasible for a group* $S \subset N$ if

$$\overline{u}(S) \overset{\text{def}}{=} \sum_{i \in S} \overline{u}^i \leq \sum_{k=1}^{K} v(S^k) \tag{1}$$

for some partition $\{S^1, \ldots, S^K\}$ of S.

Given $\varepsilon \geq 0$, a payoff vector $\overline{u} \in \mathbb{R}^N$ is in the *weak ε-core* of the game (N, v) if it is feasible and if there is a group of players $N^0 \subset N$ such that

$$\frac{|N \backslash N^0|}{|N|} \leq \varepsilon \tag{2}$$

and, for all groups $S \subset N^0$,

$$\overline{u}(S) \geq v(S) - \varepsilon |S| \tag{3}$$

where $|S|$ is the cardinality of the set S. (It would be possible to use two different values for epsilon in expressions (2) and (3). For simplicity, we have chosen to take the same value for epsilon in both expressions.) A payoff vector $\overline{u}$ is in the *uniform ε-core* (or simply in the *ε-core*) if if is feasible and if (3) holds for *all* groups $S \subset N$. When $\varepsilon = 0$, then both notions of ε-cores will be called simply the *core*.

Example 1 (continued) The glove game (N, v) described in Example 1 has the happy feature that the core is always nonempty. For the game to be of interest, we will suppose that there is least one player of each type (that is, there is at least one player with a RH glove and one player with a LH glove). If $|N_1| = |N_2|$ any payoff vector assigning the same share of a dollar to each player with a LH glove and the remaining share of a dollar to each player with a RH glove is in the core. If there are more players of one type, say $|N_1| > |N_2|$ for specificity, then any payoff vector in the core assigns \$1 to each player of the scarce type; that is, players with a RH glove each receive 0 while players with a LH glove each receive \$1.

Not all games have nonempty cores, as the following example illustrates.

Example 2 (A simple majority game with an empty core) Let $N = \{1, 2, 3\}$ and define the function v as follows:

$$v(S) = \begin{cases} 0 \text{ if } |S| = 1, \\ 1 \text{ otherwise}. \end{cases}$$

It is easy to see that the core of the game is empty. For if a payoff vector $\overline{u}$ were in the core, then it must hold that for any $i \in N$, $\overline{u}_i \geq 0$ and for any $i, j \in N, \overline{u}_i + \overline{u}_j \geq 1$. Moreover, feasibility dictates that $\overline{u}_1 + \overline{u}_2 + \overline{u}_3 \leq 1$. This is impossible; thus, the core is empty.

Before leaving this example, let us ask whether it would be possible to subsidize the players by increasing the payoff to the total player set N and, by doing so, ensure that the core of the game with a subsidy is nonempty. We leave it to the reader to verify that if $v(N)$ were increased to \$3/2 (or more), the new game would have a nonempty core.

Let (N, v) be a game and let $i, j \in N$. Then players i and j are *substitutes* if, for all groups $S \subset N$ with $i, j \notin S$ it holds that

$$v(S \cup \{i\}) = v(S \cup \{j\}).$$

Let (N, v) be a game and let $\overline{u} \in \mathbb{R}^N$ be a payoff vector for the game. If for all players i and j who are substitutes it holds that $\overline{u}_i = \overline{u}_j$ then $\overline{u}$ has the *equal treatment property*. Note that if there is a partition of N into T subsets, say $N_1, \ldots, N_T$, where all players in each subset N_t are substitutes for each other, then we can *represent* $\overline{u}$ by a vector $\overline{\overline{u}} \in \mathbb{R}^T$ where, for each t, it holds that $\overline{\overline{u}}_t = \overline{u}_i$ for all $i \in N_t$.

Essential Superadditivity

We wish to treat games where the worth of a group of players is independent of the total player set in which it is embedded and an option open to the members of a group is to partition themselves into smaller groups; that is, we treat games that are *essentially superadditive*. This is built into our the definition of feasibility above, (1). An alternative approach, which would still allow us to treat situations where it is optimal for players to form groups smaller than the total player set, would be to assume that v is the "superadditive cover" of some other worth function v'. Given a not-necessarily-superadditive function v', for each group S define $v(S)$ by:

$$v(S) = \max \sum v'(S^k) \tag{4}$$

where the maximum is taken over all partitions $\{S^k\}$ of S; the function v is the *superadditive cover* of v'. Then the notion of feasibility requiring that a payoff vector $\overline{u}$ is feasible only if

$$\overline{u}(N) \leq v(N)\,, \tag{5}$$

gives an equivalent set of feasible payoff vectors to those of the game (N, v') with the definition of feasibility given by (1).

The following Proposition may be well known and is easily proven. This result was already well understood in Gillies [27] and applications have appeared in a number of papers in the theoretical literature of game theory; see, for example (for $\varepsilon = 0$) Aumann and Dreze [6] and Kaneko and Wooders [33]. It is also well known in club theory and the theory of economies with many players and local public goods.

Proposition 1 *Given $\varepsilon \geq 0$, let (N, v') be a game. A payoff vector $\overline{u} \in R^N$ is in the weak, respectively uniform, ε-core of (N, v') if and only if it is in the weak, respectively uniform, ε-core of the superadditive cover game, say (N, v), where v is defined by (4).*

A Market

In this section we introduce the definition, from Shapley and Shubik [60], of a market. Unlike Shapley and Shubik, however, we do not assume concavity of utility functions. A *market* is taken to be an economy where all participants have continuous utility functions over a finite set of commodities that are all linear in one commodity, thought of as an "idealized" money. Money can be consumed in any amount, possibly negative. For later convenience we will consider an economy where there is a finite set of types of participants in the economy and all participants of the same type have the same endowments and preferences.

Consider an economy with $T + 1$ types of commodities. Denote the set of participants by

$$N = \{(t, q) : t = 1, \ldots, T, \quad \text{and } q = 1, \ldots, n_t\}\,.$$

Assume that all participants of the same type, (t, q), $q = 1, \ldots, n_t$ have the same utility functions given by

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi$$

where $y \in \mathbb{R}^T_+$ and $\xi \in \mathbb{R}$. Let $a^{tq} \in \mathbb{R}^T_+$ be the *endowment* of the (t, q)th player of the first T commodities. The *total endowment* is given by $\sum_{(t,q)\in N} a^{tq}$. For simplicity and without loss of generality, we can assume that no participant is endowed with any nonzero amount of the $(T + 1)^{th}$ good, the "money" or medium of exchange. One might think of utilities as being measured in money. It is because of the transferability of money that utilities are called "transferable".

Remark 1 Instead of assuming that money can be consumed in negative amounts one might assume that endowments of money are sufficiently large so that no equilibrium allocates any participant a negative amount of money. For further discussion of transferable utility see, for example, Bergstrom and Varian [9] or Kaneko and Wooders [34] .

Given a group $S \subset N$, a *S-allocation of commodities* is a set

$$\left\{ (y^{tq}, \xi^{tq}) \in \mathbb{R}^T_+ \times \mathbb{R} : \sum_{(t,q)\in S} y^{tq} \leq \sum_{(t,q)\in S} a^{tq} \text{ and } \sum_{(t,q)\in S} \xi^{tq} \leq 0 \right\} ;$$

that is, a S-allocation is a redistribution of the commodities owned by the members of S among themselves and monetary transfers adding up to no more than zero. When $S = N$, a S-allocation is called simply an *allocation*.

With the price of the $(T + 1)^{th}$ commodity ξ set equal to 1, a *competitive outcome* is a price vector p in $\mathbb{R}^T$, listing prices for the first T commodities, and an allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}^T \times \mathbb{R} : (t, q) \in N\}$ for which

$$\begin{aligned}
&\text{(a) } u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) \geq u_t(\widehat{y}) - p \cdot (\widehat{y} - a^{tq}) \\
&\quad \text{for all } \widehat{y} \in \mathbb{R}^T_+, (t, q) \in N\,, \\
&\text{(b) } \textstyle\sum_{(t,q)\in N} y^{tq} = \sum_{(t,q)} a^{tq} = \overline{y}\,, \\
&\text{(c) } \xi^{tq} = p \cdot (y^{tq} - a^{tq}) \quad \text{for all} \quad (t, q) \in N \quad \text{and} \\
&\text{(d) } \textstyle\sum_{(t,q)\in N} \xi^{tq} = 0\,.
\end{aligned} \tag{6}$$

Given a competitive outcome with allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}^T_+ \times \mathbb{R} : (t, q) \in N\}$ and price vector p, the *competitive payoff to the $(t, q)^{th}$ participant* is $u(y^{tq}) - p \cdot (y^{tq} - a^{tq})$. A *competitive payoff vector* is given by

$$(u(y^{tq}) - p \cdot (y^{tq} - a^{tq}) : (t, q) \in N)\,.$$

In the following we will assume that for each t, all participants of type t have the same endowment; that is, for each t, it holds that $a^{tq} = a^{tq'}$ for all $q, q' = 1, \ldots, n_t$. In this case, every competitive payoff has the equal treatment property;

$$u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) = u_t(y^{tq'}) - p \cdot (y^{tq'} - a^{tq'})$$

for all q, q' and for each t. It follows that a competitive payoff vector can be represented by a vector in $\mathbb{R}^T$ with one component for each player type.

It is easy to generate a game from the data of an economy. For each group of participants $S \subset N$, define

$$v(S) = \max \sum_{tq \in S} u_t(y^{tq}, \xi^{tq})$$

where the maximum is taken over the set of S-allocations. Let (N, v) denote a game derived from a market.

Under the assumption of concavity of the utility functions of the participants in an economy, Shapley and Shubik [60] show that a competitive outcome for the market exists and that the competitive payoff vectors are in the core of the game. (Since [22], such results have been obtained in substantially more general models of economies.)

Market-Game Equivalence

To facilitate exposition of the theory of games with many players and the equivalence of markets and games, we consider games derived from a common underlying structure and with a fixed number of types of players, where all players of the same type are substitutes for each other.

Pregames

Let T be a positive integer, to be interpreted as a number of player types. A *profile* $s = (s_1, \ldots, s_T) \in \mathbf{Z}_+^T$, where $\mathbf{Z}_+^T$ is the T-fold Cartesian product of the non-negative integers $\mathbf{Z}_+$, describes a group of players by the numbers of players of each type in the group. Given profile s, define the *norm* or *size* of s by

$$\|s\| \overset{\text{def}}{=} \sum_t s_t \, ,$$

simply the total number of players in a group of players described by s. A *subprofile of a profile* $n \in \mathbf{Z}_+^T$ is a profile s satisfying $s \leq n$. A *partition of a profile* s is a collection of subprofiles $\{s^k\}$ of n, not all necessarily distinct, satisfying

$$\sum_k s^k = s \, .$$

A partition of a profile is analogous to a partition of a set except that all members of a partition of a set are distinct.

Let Ψ be a function from the set of profiles $\mathbf{Z}_+^T$ to $\mathbb{R}_+$ with $\Psi(0) = 0$. The value $\Psi(s)$ is interpreted as the total payoff a group of players with profile s can achieve from collective activities of the group membership and is called the *worth of the profile* s.

Given Ψ, define a worth function Ψ^*, called the *superadditive cover* of Ψ, by

$$\Psi^*(s) \overset{\text{def}}{=} \max \sum_k \Psi(s^k) \, ,$$

where the maximum is taken over the set of all partitions $\{s^k\}$ of s . The function Ψ is said to be *superadditive* if the worth functions Ψ and Ψ^* are equal.

We define a *pregame* as a pair (T, Ψ) where $\Psi : \mathbf{Z}_+^T \to \mathbb{R}_+$. As we will now discuss, a pregame can be used to generate multiple games. To generate a game from a pregame, it is only required to specify a total player set N and the numbers of players of each of T types in the set. Then the pregame can be used to assign a worth to every group of players contained in the total player set, thus creating a game.

A *game determined by the pregame* (T, Ψ), which we will typically call a *game* or a *game with side payments*, is a pair $[n; (T, \Psi)]$ where n is a profile. A *subgame* of a game $[n; (T, \Psi)]$ is a pair $[s; (T, \Psi)]$ where s is a subprofile of n.

With any game$[n; (T, \Psi)]$ we can associate a game (N, v) in the form introduced earlier as follows: Let

$$N = \{(t, q) : t = 1, \ldots, T \text{ and } q = 1, \ldots, n_t\}$$

be a *player set* for the game. For each subset $S \subset N$ define the *profile of* S, denoted by prof(S)$\in \mathbf{Z}_+^T$, by its components

$$\text{prof}(S)_t \overset{\text{def}}{=} \left|\{S \cap \{(t', q) : t' = t \text{ and } q = 1, \ldots, n_t\}\right|$$

and define

$$v(S) \overset{\text{def}}{=} \Psi(\text{prof}(S)) \, .$$

Then the pair (N, v) satisfies the usual definition of a game with side payments. For any $S \subset N$, define

$$v^*(S) \overset{\text{def}}{=} \Psi^*(\text{prof}(S)) \, .$$

The game (N, v^*) is the *superadditive cover of* (N, v).

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. For each nonempty subset S of N define

$$\overline{u}(S) \overset{\text{def}}{=} \sum_{(t,q) \in S} \overline{u}^{tq} \, .$$

A payoff vector $\overline{u}$ is *feasible for* S if

$$\overline{u}(S) \leq v^*(S) = \Psi^*(\text{prof}(S)) \, .$$

If $S = N$ we simply say that the payoff vector $\overline{u}$ is *feasible* if

$$\overline{u}(N) \leq v^*(N) = \Psi^*(\text{prof}(N)) \, .$$

Note that our definition of feasibility is consistent with essential superadditivity; a group can realize at least as large a total payoff as it can achieve in any partition of the group and one way to achieve this payoff is by partitioning into smaller groups.

A payoff vector $\overline{u}$ satisfies the *equal-treatment property* if $\overline{u}^{tq} = \overline{u}^{tq'}$ for all $q, q' \in \{1, \dots, n_t\}$ and for each $t = 1, \dots, T$.

Let $[n, (T, \Psi)]$ be a game and let β be a collection of subprofiles of n. The collection is a *balanced collection of subprofiles of* n if there are positive real numbers γ_s for $s \in \beta$ such that $\sum_{s\in\beta} \gamma_s s = n$. The numbers γ_s are called *balancing weights*. Given real number $\varepsilon \geq 0$, the game $[n; (T, \Psi)]$ is *ε-balanced* if for every balanced collection β of subprofiles of n it holds that

$$\Psi^*(n) \geq \sum_{s\in\beta} \gamma_s \left(\Psi(s) - \varepsilon\|s\|\right) \tag{7}$$

where the balancing weights for β are given by γ_s for $s \in \beta$. This definition extends that of Bondareva [13] and Shapley [56] to games with player types. Roughly, a game is (ε) balanced if allowing "part time" groups does not improve the total payoff (by more than ε per player). A game $[n; (T, \Psi)]$ is *totally balanced* if every subgame $[s; (T, \Psi)]$ is balanced.

The *balanced cover game* generated by a game $[n; (T, \Psi)]$ is a game $[n; (T, \Psi^b)]$ where

1. $\Psi^b(s) = \Psi(s)$ for all $s \neq n$ and
2. $\Psi^b(n) \geq \Psi(n)$ and $\Psi^b(n)$ is as small as possible consistent with the nonemptiness of the core of $[n; (T, \Psi^b)]$.

From the Bondareva–Shapley Theorem it follows that $\Psi^b(n) = \Psi^*(n)$ if and only if the game $[n; (T, \Psi)]$ is balanced (ε-balanced, with $\varepsilon = 0$).

For later convenience, the notion of the balanced cover of a pregame is introduced. Let (T, Ψ) be a pregame. For each profile s, define

$$\Psi^b(s) \overset{\text{def}}{=} \max_{\beta} \sum_{g\in\beta} \gamma_g \Psi(g) \,, \tag{8}$$

where the maximum is taken over all balanced collections β of subprofiles of s with weights γ_g for $g \in \beta$. The pair (T, Ψ^b) is called the *balanced cover pregame* of (T, Ψ). Since a partition of a profile is a balanced collection it is immediately clear that $\Psi^b(s) \geq \Psi^*(s)$ for every profile s.

Premarkets

In this section, we introduce the concept of a premarket and re-state results from Shapley and Shubik [60] in the context of pregames and premarkets.

Let $L + 1$ be a number of types of commodities and let $\{\widehat{u}_t(y, \xi) : t = 1, \dots, T\}$ denote a finite number of functions, called *utility functions*, of the form

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi \,,$$

where $y \in \mathbb{R}^L_+$ and $\xi \in \mathbb{R}$. (Such functions, in the literature of economics, are commonly called *quasi-linear*). Let $\{a^t \in \mathbb{R}^L_+ : t = 1, \dots, T\}$ be interpreted as a set of *endowments*. We assume that $u_t(a^t) \geq 0$ for each t. For $t = 1, \dots, T$ we define $c^t \overset{\text{def}}{=} (u_t(\cdot), a^t)$ as a *participant type* and let $\mathbb{C} = \{c^t : t = 1, \dots, T\}$ be the set of participant types. Observe that from the data given by $\mathbb{C}$ we can construct a market by specifying a set of participants N and a function from N to $\mathbb{C}$ assigning endowments and utility functions – types – to each participant in N. A *premarket* is a pair $(T, \mathbb{C})$.

Let $(T, \mathbb{C})$ be a premarket and let $s = (s_1, \dots, s_T) \in \mathbf{Z}^T_+$. We interpret s as representing a group of economic participants with s_t participants having utility functions and endowments given by c^t for $t = 1, \dots, T$; for each t, that is, there are s_t participants in the group with type c^t. Observe that the data of a premarket gives us sufficient data to generate a pregame. In particular, given a profile $s = (s_1, \dots, s_T)$ listing numbers of participants of each of T types, define

$$W(s) \overset{\text{def}}{=} \max \sum_t s_t u_t(y^t)$$

where the maximum is taken over the set $\{y^t \in \mathbb{R}^L_+ : t = 1, \dots, T \text{ and } \sum_t s_t y^t = \sum_t a^t y^t\}$. Then the pair (T, W) is a *pregame generated by the premarket*.

The following Theorem is an extension to premarkets or a restatement of a result due to Shapley and Shubik [60].

Theorem 1 *Let $(T, \mathbb{C})$ be a premarket derived from economic data in which all utility functions are concave. Then the pregame generated by the premarket is totally balanced.*

Direct Markets and Market-Game Equivalence

Shapley and Shubik [60] introduced the notion of a direct market derived from a totally balanced game. In the direct market, each player is endowed with one unit of a commodity (himself) and all players in the economy have the same utility function. In interpretation, we might think of this as a labor market or as a market for productive factors, (as in [50], for example) where each player owns one unit of a commodity. For games with player types as in this essay, we take the player types of the game as the commodity types of a market and assign all players in the market the same utility function, derived from the worth function of the game.

Let (T, Ψ) be a pregame and let $[n; (T, \Psi)]$ be a derived game. Let $N = \{(t, q) : t = 1, \dots, T$ and $q = 1, \dots, n_t$ for each $t\}$ denote the set of players in the game where all participants $\{(t', q) : q = 1, \dots, n_{t'}\}$ are of type t' for each $t' = 1, \dots, T$. To construct the direct market generated by a derived game $[n; (T, \Psi)]$, we take the commodity space as $\mathbb{R}^T_+$ and suppose that each participant in the market of type t is endowed with one unit of the tth commodity, and thus has endowment $\mathbf{1}_t = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^T_+$ where "1" is in the tth position. The total endowment of the economy is then given by $\sum n_t \mathbf{1}_t = n$.

For any vector $y \in \mathbb{R}^T_+$ define

$$u(y) \overset{\text{def}}{=} \max \sum_{s \leq n} \gamma_s \Psi(s) \ , \tag{9}$$

the maximum running over all $\{\gamma_s \geq 0 : s \in \mathbf{Z}^T_+, \ s \leq n\}$ satisfying

$$\sum_{s \leq n} \gamma_s s = y \ . \tag{10}$$

As noted by Shapley and Shubik [60], but for our types case, it can be verified that the function u is concave and one-homogeneous. This does not depend on the balancedness of the game $[n; (T, \Psi)]$. Indeed, one may think of u as the "balanced cover of $[n; (T, \Psi)]$ extended to $\mathbb{R}^T_+$". Note also that u is superadditive, independent of whether the pregame (T, Ψ) is superadditive. We leave it to the interested reader to verify that if Ψ were not necessarily superadditive and Ψ^* is the superadditive cover of Ψ then it holds that $\max \sum_{s \leq n} \gamma_s \Psi(s) = \max \sum_{s \leq n} \gamma_s \Psi^*(s)$.

Taking the utility function u as the utility function of each player $(t, q) \in N$ where N is now interpreted as the set of participants in a market, we have generated a market, called the *direct market*, denoted by $[n, u; (T, \Psi)]$, from the game $[n; (T, \Psi)]$.

Again, the following extends a result of Shapley and Shubik [60] to pregames.

Theorem 2 *Let $[n, u; (T, \Psi)]$ denote the direct market generated by a game $[n; (T, \Psi)]$ and let $[n; (T, u)]$ denote the game derived from the direct market. Then, if $[n; (T, \Psi)]$ is a totally balanced game, it holds that $[n; (T, u)]$ and $[n; (T, \Psi)]$ are identical.*

Remark 2 If the game $[n; (T, \Psi)]$ and every subgame $[s, (T, \Psi)]$ has a nonempty core – that is, if the game is 'totally balanced'– then the game $[n; (T, u)]$ generated by the direct market is the initially given game $[n; (T, \Psi)]$. If however the game $[n; (T, \Psi)]$ is not totally balanced then $u(s) \geq \Psi(s)$ for all profiles $s \leq n$. But, whether or not $[n; (T, \Psi)]$ is totally balanced, the game $[n; (T, u)]$ is totally balanced and coincides with the totally balanced cover of $[n; (T, \Psi)]$.

Remark 3 Another approach to the equivalence of markets and games is taken by Garratt and Qin [26], who define a class of direct lottery markets. While a player can participate in only one coalition, both ownership of coalitions and participation in coalitions is determined randomly. Each player is endowed with one unit of probability, his own participation. Players can trade their endowments at market prices. The core of the game is equivalent to the equilibrium of the direct market lottery.

Equivalence of Markets and Games with Many Players

The requirement of Shapley and Shubik [60] that utility functions be concave is restrictive. It rules out, for example situations such as economies with indivisible commodities. It also rules out club economies; for a given club structure of the set of players – in the simplest case, a partition of the total player set into groups where collective activities only occur within these groups – it may be that utility functions are concave over the set of alternatives available within each club, but utility functions need not be concave over all possible club structures. This rules out many examples; we provide a simple one below.

To obtain the result that with many players, games derived from pregames are market games, we need some further assumption on pregames. If there are many substitutes for each player, then the simple condition that *per capita payoffs are bounded* – that is, given a pregame (T, Ψ), that there exists some constant K such that $\frac{\Psi(s)}{\|s\|} < K$ for all profiles s – suffices. If, however, there may be 'scarce types', that is, players of some type(s) become negligible in the population, then a stronger assumption of 'small group effectiveness' is required. We discuss these two conditions in the next section.

Small Group Effectiveness and Per Capita Boundedness

This section discusses conditions limiting gains to group size and their relationships. This definition was introduced in Wooders [83], for NTU, as well as TU, games.

PCB A pregame (T, Ψ) satisfies *per capita boundedness* (PCB) if

$$PCB: \quad \sup_{s \in \mathbf{Z}_+^T} \frac{\Psi(s)}{\|s\|} \text{ is finite} \tag{11}$$

or equivalently,

$$\sup_{s \in \mathbf{Z}_+^T} \frac{\Psi^*(s)}{\|s\|} \text{ is finite} .$$

It is known that under the apparently mild conditions of PCB and essential superadditivity, in general games with many players of each of a finite number of player types and a fixed distribution of player types have nonempty approximate cores; Wooders [81,83]. (Forms of these assumptions were subsequently also used in Shubik and Wooders [69,70]; Kaneko and Wooders [35]; and Wooders [89,91] among others.) Moreover, under the same conditions, approximate cores have the property that most players of the same type are treated approximately equally ([81,94]; see also Shubik and Wooders [69]). These results, however, either require some assumption ruling out 'scarce types' of players, for example, situations where there are only a few players of some particular type and these players can have great effects on total feasible payoffs. Following are two examples. The first illustrates that PCB does not control limiting properties of the per capita payoff function when some player types are scarce.

Example 3 ([94]) Let $T = 2$ and let (T, Ψ) be the pregame given by

$$\Psi(s_1, s_2) = \begin{cases} s_1 + s_2 & \text{when } s_1 > 0 \\ 0 & \text{otherwise} . \end{cases}$$

The function Ψ obviously satisfies PCB. But there is a problem in defining $\lim \Psi(s_1, s_2)/s_1 + s_2$ as $s_1 + s_2$ tends to infinity, since the limit depends on how it is approached. Consider the sequence (s_1^ν, s_2^ν) where $(s_1^\nu, s_2^\nu) = (0, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 0$. Now suppose in contrast that $(s_1^\nu, s_2^\nu) = (1, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 1$. This illustrates why, to obtain the result that games with many players are market games either it must be required that there are no scarce types or some some assumption limiting the effects of scarce types must be made. We return to this example in the next section.

The next example illustrates that, with only PCB, uniform approximate cores of games with many players derived from pregames may be empty.

Example 4 ([94]) Consider a pregame (T, Ψ) where $T = \{1, 2\}$ and Ψ is the superadditive cover of the function Ψ' defined by:

$$\Psi'(s) \overset{\text{def}}{=} \begin{cases} |s| & \text{if } s_1 = 2 , \\ 0 & \text{otherwise} . \end{cases}$$

Thus, if a profile$s = (s_1, s_2)$ has $s_1 = 2$ then the worth of the profile according to Ψ' is equal to the total number of players it represents, $s_1 + s_2$, while all other profiles s have worth of zero. In the superadditive cover game the worth of a profile s is 0 if $s_1 < 2$ and otherwise is equal to s_2 plus the largest even number less than or equal to s_1.

Now consider a sequence of profiles $(s^\nu)_\nu$ where $s_1^\nu = 3$ and $s_2^\nu = \nu$ for all ν. Given $\varepsilon > 0$, for all sufficiently large player sets the uniform ε-core is empty. Take, for example, $\varepsilon = 1/4$. If the uniform ε-core were nonempty, it would have to contain an equal-treatment payoff vector.[1] For the purpose of demonstrating a contradiction, suppose that $u^\nu = (u_1^\nu, u_2^\nu)$ represents an equal treatment payoff vector in the uniform ε-core of $[s^\nu; (T, \Psi)]$. The following inequalities must hold:

$$\begin{aligned} 3u_1^\nu + \nu u_2^\nu &\le \nu + 2 , \\ 2u_1^\nu + \nu u_2^\nu &\ge \nu + 2, \text{ and} \\ u_1^\nu &\ge \tfrac{3}{4} . \end{aligned}$$

which is impossible. A payoff vector which assigns each player zero is, however, in the weak ε-core for any $\varepsilon > \frac{1}{\nu+3}$. But it is not very appealing, in situations such as this, to ignore a relatively small group of players (in this case, the players of type 1) who can have a large effect on per capita payoffs. This leads us to the next concept.

To treat the scarce types problem, Wooders [88,89,90] introduced the condition of small group effectiveness (SGE). SGE is appealing technically since it resolves the scarce types problem. It is also economically intuitive and appealing; the condition defines a class of economies that, when there are many players, generate competitive markets. Informally, SGE dictates that *almost all* gains to collective activities can be realized by relatively small groups of players. Thus, SGE is exactly the sort of assumption required to ensure that multiple, relatively small coalitions, firms, jurisdictions, or clubs, for example, are optimal or near-optimal in large economies.

[1] It is well known and easily demonstrated that the uniform ε-core of a TU game is nonempty if and only if it contains an equal treatment payoff vector. This follows from the fact that the uniform ε-core is a convex set.

A pregame (T, Ψ) satisfies *small group effectiveness, SGE*, if:

$$SGE: \quad \begin{array}{c} \text{For each real number } \varepsilon > 0, \\ \text{there is an integer } \eta_0(\varepsilon) \\ \text{such that for each profile } s, \\ \text{for some partition } \{s^k\} \text{ of } s \text{ with} \\ \|s^k\| \leq \eta_0(\varepsilon) \text{ for each subprofile } s^k, \text{ it holds that} \\ \Psi^*(s) - \sum_k \Psi(s^k) \leq \varepsilon\|s\| \, ; \end{array} \tag{12}$$

given $\varepsilon > 0$ there is a group size $\eta_0(\varepsilon)$ such that the loss from restricting collective activities within groups to groups containing fewer that $\eta_0(\varepsilon)$ members is at most ε per capita [88].[2]

SGE also has the desirable feature that if there are no 'scarce types' – types of players that appear in vanishingly small proportions– then SGE and PCB are equivalent.

Theorem 3 ([91] With 'thickness,' SGE = PCB) *(1) Let (T, Ψ) be a pregame satisfying SGE. Then the pregame satisfies PCB.*

(2) Let (T, Ψ) be a pregame satisfying PCB. Then given any positive real number ρ, construct a new pregame (T, Ψ_ρ) where the domain of Ψ_ρ is restricted to profiles s where, for each $t = 1, \ldots, T$, either $\frac{s_t}{\|s\|} > \rho$ or $s_t = 0$. Then (T, Ψ_ρ) satisfies SGE on its domain.

It can also be shown that small groups are effective for the attainment of nearly all feasible outcomes, as in the above definition, if and only if small groups are effective for improvement – any payoff vector that can be significantly improved upon can be improved upon by a small group (see Proposition 3.8 in [89]).

Remark 4 Under a stronger condition of *strict* small group effectiveness, which dictates that $\eta(\varepsilon)$ in the definition of small group effectiveness can be chosen independently of ε, stronger results can be obtained than those presented in this section and the next. We refer to Winter and Wooders [80] for a treatment of this case.

Remark 5 (On the importance of taking into account scarce types) Recall the quotation from von Neumann and Morgenstern and the discussion following the quotation. The assumption of per capita boundedness has significant consequences but is quite innocuous – ruling out the possibility of average utilities becoming infinite as economies grow large does not seem restrictive. But with only per capita boundedness, even the formation of small coalitions can have significant impacts on aggregate outcomes. With small group effectiveness, however, there is no problem of either large or small coalitions acting together – large coalitions cannot do significantly better then relatively small coalitions.

Roughly, the property of large games we next introduce is that relatively small groups of players make only "asymptotic negligible" contributions to per-capita payoffs of large groups. A pregame (Ω, Ψ) satisfies *asymptotic negligibility* if, for any sequence of profiles $\{f^\nu\}$ where

$$\begin{aligned} &\|f^\nu\| \to \infty \text{ as } \nu \to \infty, \\ &\sigma(f^\nu) = \sigma(f^{\nu'}) \quad \text{for all} \quad \nu \quad \text{and} \quad \nu' \quad \text{and} \\ &\lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \quad \text{exists ,} \end{aligned} \tag{13}$$

then for any sequence of profiles $\{\ell^\nu\}$ with

$$\lim_{\nu\to\infty} \frac{\|\ell^\nu\|}{\|f^\nu\|} = 0 \, , \tag{14}$$

it holds that

$$\begin{aligned} &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu + \ell^\nu\|}{\|f^\nu + \ell^\nu\|} \quad \text{exists, and} \\ &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu + \ell^\nu\|}{\|f^\nu + \ell^\nu\|} = \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \, . \end{aligned} \tag{15}$$

Theorem 4 ([89,95]) *A pregame (T, Ψ) satisfies SGE if and only if it satisfies PCB and asymptotic negligibility*

Intuitively, asymptotic negligibility ensures that vanishingly small percentages of players have vanishingly small effects on aggregate per-capita worths. It may seem paradoxical that SGE, which highlights the importance of relatively small groups, is equivalent to asymptotic negligibility. To gain some intuition, however, think of a marriage model where only two-person marriages are allowed. Obviously two-person groups are (strictly) effective, but also, in large player sets, no two persons can have a substantial affect on aggregate per-capita payoffs.

Remark 6 Without some assumptions ensuring essential superadditivity, at least as incorporated into our definition of feasibility, nonemptiness of approximate cores of large games cannot be expected; superadditivity assumptions (or the close relative, essential superadditivity) are heavily relied upon in all papers on large games cited. In the context of economies, superadditivity is a sort of monotonicity of preferences or production functions assumption, that is, superadditivity of Ψ implies that for all $s, s' \in \mathbf{Z}_+^T$, it holds that $\Psi(s + s') \geq \Psi(s) + \Psi(s')$. Our assumption of small group effectiveness, SGE, admits non-monotonicities. For example, suppose that 'two is company, three or more is a crowd,' by supposing there is only one commodity and

[2] Exactly the same definition applies to situations with a compact metric space of player types, c.f. Wooders [84,88].

by setting $\Psi(2) = 2$, $\Psi(n) = 0$ for $n \neq 2$. The reader can verify, however, that this example satisfies small group effectiveness since $\Psi^*(n) = n$ if n is even and $\Psi^*(n) = n{-}1$ otherwise. Within the context of pregames, requiring the superadditive cover payoff to be approximately realizable by partitions of the total player set into relatively small groups is the weakest form of superadditivity required for the equivalence of games with many players and concave markets.

Derivation of Markets from Pregames Satisfying SGE

With SGE and PCB in hand, we can now derive a premarket from a pregame and relate these concepts.

To construct a limiting direct premarket from a pregame, we first define an appropriate utility function. Let (T, Ψ) be a pregame satisfying SGE. For each vector x in $\mathbb{R}^T_+$ define

$$U(x) \overset{\text{def}}{=} \|x\| \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \tag{16}$$

where the sequence $\{f^\nu\}$ satisfies

$$\begin{aligned} &\lim\nolimits_{\nu\to\infty} \tfrac{f^\nu}{\|f^\nu\|} = \tfrac{x}{\|x\|} \\ &\text{and} \\ &\|f^\nu\| \to \infty\,. \end{aligned} \tag{17}$$

Theorem 5 ([84,91]) *Assume the pregame (T, Ψ) satisfies small group effectiveness. Then for any $x \in \mathbb{R}^T_+$ the limit (16) exists. Moreover, $U(\cdot)$ is well-defined, concave and 1-homogeneous and the convergence is uniform in the sense that, given $\varepsilon > 0$ there is an integer η such that for all profiles s with $\|s\| \leq \eta$ it holds that*

$$\left| U\left(\frac{s}{\|s\|}\right) - \frac{\Psi^*(s)}{\|s\|} \right| \leq \varepsilon\,.$$

From Wooders [91] (Theorem 4), if arbitrarily small percentages of players of any type that appears in games generated by the pregame are ruled out, then the above result holds under per capita boundedness [91] (Theorem 6). As noted in the introduction to this paper, for the TU case, the concavity of the limiting utility function, for the model of Wooders [83] was first noted by Aumann [5]. The concavity is shown to hold with a compact metric space of player types in Wooders [84] and is simplified to the finite types case in Wooders [91].

Theorem 5 follows from the facts that the function U is superadditive and 1-homogeneous on its domain. Since U is concave, it is continuous on the interior of its domain; this follows from PCB. Small group effectiveness ensures that the function U is continuous on its entire domain [91](Lemma 2).

Theorem 6 ([91]) *Let (T, Ψ) be a pregame satisfying small group effectiveness and let (T, U) denote the derived direct market pregame. Then (T, U) is a totally balanced market game. Moreover, U is one-homogeneous, that is, $U(\lambda x) = \lambda U(x)$ for any non-negative real number λ.*

In interpretation, T denotes a number of types of players/commodities and U denotes a utility function on $\mathbb{R}^T_+$. Observe that when U is restricted to profiles (in $\mathbf{Z}^T_+$), the pair (T, U) is a pregame with the property that every game $[n; (T, U)]$ has a nonempty core; thus, we will call (T, U) the *premarket generated by the pregame* (T, Ψ). That every game derived from (T, U) has a nonempty core is a consequence of the Shapley and Shubik [60] result that market games derived from markets with concave utility functions are totally balanced.

It is interesting to note that, as discussed in Wooders (Section 6 in [91]), if we restrict the number of commodities to equal the number of player types, then the utility function U is *uniquely* determined. (If one allowed more commodities then one would effectively have 'redundant assets'.) In contrast, for games and markets of fixed, finite size, as demonstrated in Shapley and Shubik [62], even if we restrict the number of commodities to equal the number of player types, given any nonempty, compact, convex subset of payoff vectors in the core, it is possible to construct utility functions so that this subset coincides with the set of competitive payoffs. Thus, in the Shapley and Shubik approach, equivalence of the core and the set of price-taking competitive outcomes for the direct market is only an artifact of the method used there of constructing utility functions from the data of a game and is quite distinct from the equivalence of the core and the set of competitive payoff vectors as it is usually understood (that is, in the sense of Debreu and Scarf [22] and Aumann [4]. See also Kalai and Zemel [31,32] which characterize the core in multi-commodity flow games.

Cores and Approximate Cores

The concept of the core clearly was important in the work of Shapley and Shubik [59,60,62] and is also important for the equivalence of games with many players and market games. Thus, we discuss the related results of nonemptiness of approximate cores and convergence of approximate cores to the core of the 'limit' – the game where all players have utility functions derived from a pregame and large numbers of players. First, some terminology is required. A vector p is a *subgradient at* x of the concave function U if $U(y) - U(x) \leq p \cdot (y - x)$ for all y. One might think of a subgradient as a bounding hyperplane. To avoid any confusion it might be helpful to note that, as Mas-Colell [46] remarks: " Strictly speaking, one should use the term *subgradient* for convex functions and *supergradient* for concave. But this is cumbersome", (p. 29–30 in [46]).

For ease of notation, equal-treatment payoff vectors for a game $[n; (T, \Psi)]$ will typically be represented as vectors in $\mathbb{R}^T$. An *equal-treatment payoff vector*, or simply a *payoff vector* when the meaning is clear, is a point $\overline{x}$ in $\mathbb{R}^T$. The t^{th} component of $\overline{x}$, $\overline{x}_t$, is interpreted as the payoff to each player of type t. The feasibility of an equal-treatment payoff vector $\overline{x} \in \mathbb{R}^T$ for the game $[n; (T, \Psi)]$ can be expressed as:

$$\Psi^*(n) \geq \overline{x} \cdot n \, .$$

Let $[n; (T, \Psi)]$ be a game determined by a pregame (T, Ψ), let ε be a non-negative real number, and let $\overline{x} \in \mathbb{R}^T$ be a (equal-treatment) payoff vector. Then $\overline{x}$ is in the *equal-treatment* ε*-core* of $[n; (T, \Psi)]$ or simply "in the ε-core" when the meaning is clear, if $\overline{x}$ is feasible for $[n; (T, \Psi)]$ and

$$\Psi(s) \leq \overline{x} \cdot s + \varepsilon \|s\| \text{ for all subprofiles } s \text{ of } n \, .$$

Thus, the equal-treatment ε-core is the set

$$\begin{aligned} C(n; \varepsilon) \overset{\text{def}}{=} \{&\overline{x} \in \mathbb{R}^T_+ : \Psi^*(n) \geq \overline{x} \cdot n \quad \text{and} \\ &\Psi(s) \leq \overline{x} \cdot s + \varepsilon \|s\| \quad \text{for all subprofiles} \quad s \text{ of } n\} \, . \end{aligned} \tag{18}$$

It is well known that the ε-core of a game with transferable utility is nonempty if and only if the equal-treatment ε-core is nonempty.

Continuing with the notation above, for any $s \in \mathbb{R}^T_+$, let $\Pi(s)$ denote the set of subgradients to the function U at the point s;

$$\Pi(s) \overset{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot s = U(s) \text{ and } \pi \cdot s' \geq U(s') \text{ for all } s' \in \mathbb{R}^T_+\} \, . \tag{19}$$

The elements in $\Pi(s)$ can be interpreted as equal-treatment core payoffs to a limiting game with the mass of players of type t given by s_t. The core payoff to a player is simply the value of the one unit of a commodity (himself and all his attributes, including endowments of resources) that he owns in the direct market generated by a game. Thus $\Pi(\cdot)$ is called *the limiting core correspondence* for the pregame (T, Ψ). Of course $\Pi(\cdot)$ is also the limiting core correspondence for the pregame (T, U).

Let $\widehat{\Pi}(n) \subset \mathbb{R}^T$ denote *equal-treatment core of the market game* $[n; (T, u)]$:

$$\widehat{\Pi}(n) \overset{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot n = u(n) \text{ and } \pi \cdot s \geq u(s) \text{ for all } s \in \mathbf{Z}^T_+, s \leq n\} \, . \tag{20}$$

Given any player profile n and derived games $[n; (T, \Psi)]$ and $[n; (T, U)]$ it is interesting to observe the distinction between the equal-treatment core of the game $[n; (T, U)]$, denoted by $\widehat{\Pi}(n)$, defined by (20), and the set $\Pi(n)$ (that is, $\Pi(\overline{x})$ with $\overline{x} = n$). The definitions of $\Pi(n)$ and $\widehat{\Pi}(n)$ are the same except that the qualification "$s \leq n$" in the definition of $\widehat{\Pi}(n)$ does not appear in the definition of $\Pi(n)$. Since $\Pi(n)$ is the *limiting* core correspondence, it takes into account arbitrarily large coalitions. For this reason, for any $\overline{x} \in \Pi(n)$ and $\widehat{x} \in \widehat{\Pi}(n)$ it holds that $\overline{x} \cdot n \geq \widehat{x} \cdot n$. A simple example may be informative.

Example 5 Let (T, Ψ) be a pregame where $T = 1$ and $\Psi(n) = n - \frac{1}{n}$ for each $n \in \mathbb{Z}_+$, and let $[n; (T, \Psi)]$ be a derived game. Then $\Pi(n) = \{1\}$ while $\widehat{\Pi}(n) = \{(1 - \frac{1}{n^2})\}$.

The following Theorem extends a result due to Shapley and Shubik [62] stated for games derived from pregames.

Theorem 7 ([62]) *Let* $[n; (T, \Psi)]$ *be a game derived from a pregame and let* $[n, u; (T, \Psi)]$ *be the direct market generated by* $[n; (T, \Psi)]$. *Then the equal-treatment core* $\widehat{\Pi}(n)$ *of the game* $[n; (T, u)]$ *is nonempty and coincides with the set of competitive price vectors for the direct market* $[n, u; (T, \Psi)]$.

Remark 7 Let (T, Ψ) be a pregame satisfying PCB. In the development of the theory of large games as models of competitive economies, the following function on the space of profiles plays an important role:

$$\lim_{r \to \infty} \frac{\Psi^*(rf)}{r} \, ;$$

see, for example, Wooders [81] and Shubik and Wooders [69]. For the purposes of comparison, we introduce another definition of a limiting utility function. For each vector x in $\mathbb{R}^T_+$ with rational components let $r(x)$ be the smallest integer such that $r(x)x$ is a vector of integers. Therefore, for each rational vector x, we can define

$$\hat{U}(x) \overset{\text{def}}{=} \lim_{\nu \to \infty} \frac{\Psi^*(\nu r(x) x)}{\nu r(x)} \, .$$

Since Ψ^* is superadditive and satisfies per capita boundedness, the above limit exists and $\hat{U}(\cdot)$ is well-defined. Also, $\hat{U}(x)$ has a continuous extension to any closed subset strictly in the interior of $\mathbb{R}^T_+$. The function $\hat{U}(x)$, however, may be discontinuous at the boundaries of $\mathbb{R}^T_+$. For example, suppose that $T = 2$ and

$$\Psi^*(k,n) = \begin{cases} k+n & \text{when } k > 0 \\ 0 & \text{otherwise .} \end{cases}$$

The function Ψ^* obviously satisfies PCB but does not satisfy SGE. To see the continuity problem, consider the sequences $\{x^\nu\}$ and $\{y^\nu\}$ of vectors in $\mathbb{R}^2_+$ where $x^\nu = (\frac{1}{\nu}, \frac{\nu-1}{\nu})$ and $y^\nu = (0,\nu)$. Then $\lim_{\nu\to\infty} x^\nu = \lim_{\nu\to\infty} y^\nu = (0,1)$ but $\lim_{\nu\to\infty} \hat{U}(x^\nu) = 1$ while $\lim_{\nu\to\infty} \hat{U}(y^\nu) = 0$. SGE is precisely the condition required to avoid this sort of discontinuity, ensuring that the function U is continuous on the boundaries of $\mathbb{R}^T_+$.

Before turning to the next section, let us provide some additional interpretation for $\widehat{\Pi}(n)$. Suppose a game $[n;(T,\Psi)]$ is one generated by an economy, as in Shapley and Shubik [59] or Owen [50], for example. Players of different types may have different endowments of private goods. An element π in $\widehat{\Pi}(n)$ is an equal-treatment payoff vector in the core of the balanced cover game generated by $[n;(T,\Psi)]$ and can be interpreted as listing prices for player types where π_t is the price of a player of type t; this price is a price for the player himself, *including* his endowment of private goods.

Nonemptiness and Convergence of Approximate Cores of Large Games

The next Proposition is an immediate consequence of the convergence of games to markets shown in Wooders [89,91] and can also be obtained as a consequence of Theorem 5 above.

Proposition 2 (Nonemptiness of approximate cores) *Let (T,Ψ) be a pregame satisfying SGE. Let ε be a positive real number. Then there is an integer $\eta_1(\varepsilon)$ such that any game $[n;(T,\Psi)]$ with $\|n\| \geq \eta_1(\varepsilon)$ has a nonempty uniform ε-core.*

(Note that no assumption of superadditivity is required but only because our definition of feasibility is equivalent to feasibility for superadditive covers.)

The following result was stated in Wooders [89]. For more recent results see Wooders [94].

Theorem 8 ([89] Uniform closeness of (equal-treatment) approximate cores to the core of the limit game) *Let (T,Ψ) be a pregame satisfying SGE and let $\Pi(\cdot)$ be as defined above. Let $\delta > 0$ and $\rho > 0$ be positive real numbers. Then there is a real number ε^* with $0 < \varepsilon^*$ and an integer $\eta_0(\delta,\rho,\varepsilon^*)$ with the following property: for each positive $\varepsilon \in (0,\varepsilon^*]$ and each game $[f;(T,\Psi)]$ with $\|f\| > \eta_0(\delta,\rho,\varepsilon^*)$ and $f_t/\|f\| \geq \rho$ for each $t = 1,\dots,T$, if $C(f;\varepsilon)$ is nonempty then both*

$$\operatorname{dist}[C(f;\varepsilon),\Pi(f)] < \delta \text{ and } \operatorname{dist}[C(f;\varepsilon),\widehat{\Pi}(f)] < \delta\ ,$$

where 'dist' is the Hausdorff distance with respect to the sum norm on $\mathbb{R}^T$.

Note that this result applies to games derived from diverse economies, including economies with indivisibilities, nonmonotonicities, local public goods, clubs, and so on.

Theorem 8 motivates the question of whether approximate cores of games derived from pregames satisfying small group effectiveness treat players most of the same type nearly equally. The following result, from Wooders [81,89,93] answers this question.

Theorem 9 *Let (T,Ψ) be a pregame satisfying SGE. Then given any real numbers $\gamma > 0$ and $\lambda > 0$ there is a positive real number ε^* and an integer ρ such that for each $\varepsilon \in [0, \varepsilon^*]$ and for every profile $n \in \mathbb{Z}^T_+$ with $\|n\|_1 > \rho$, if $x \in \mathbb{R}^N$ is in the uniform ε-core of the game $[n,\Psi]$ with player set*

$$N = \{(t,q) : t = 1,\dots,T \text{ and, for each } t,\ q = 1,\dots,n_t\}$$

then, for each $t \in \{1,\dots,T\}$ with $\frac{n_t}{\|n\|_1} \geq \frac{\lambda}{2}$ it holds that

$$|\{(t,q) : |x^{tq} - z_t| > \gamma\}| < \lambda n_t\}\ ,$$

where, for each $t = 1,\dots,T$,

$$z_t = \frac{1}{n_t}\sum_{q=1}^{n_t} x^{tq}\ ,$$

the average payoff received by players of type t.

Shapley Values of Games with Many Players

Let (N,ν) be a game. The *Shapley value* of a superadditive game is the payoff vector whose ith component is given by

$$SH(\nu,i) = \frac{1}{|N|}\sum_{J=0}^{|N|-1} \frac{1}{\binom{|N|-1}{J}} \sum_{\substack{S\subset N\backslash\{i\} \\ |S|=J}} \left[\nu(S\cup\{i\}) - \nu(S)\right]\ .$$

To state the next Theorem, we require one additional definition. Let (T, Ψ) be a pregame. The pregame satisfies *boundedness of marginal contributions* (BMC) if there is a constant M such that

$$|\Psi(s + 1_t) - \Psi(s)| \leq M$$

for all vectors $1_t = (0, \ldots, 0, 1_{t^{th} \text{ place}}, 0, \ldots 0)$ for each $t = 1, \ldots, T$. Informally, this condition bounds marginal contributions while SGE bounds average contributions. That BMC implies SGE is shown in Wooders [89]. The following result restricts the main Theorem of Wooders and Zame [96] to the case of a finite number of types of players.

Theorem 10 ([96]) *Let (T, Ψ) be a superadditive pregame satisfying boundedness of marginal contributions. For each $\varepsilon > 0$ there is a number $\delta(\varepsilon) > 0$ and an integer $\mu(\varepsilon)$ with the following property:*

> *If $[n, (T, \Psi)]$ is a game derived from the pregame, for which $n_t > \mu(\varepsilon)$ for each t, then the Shapley value of the game is in the (weak) ε-core.*

Similar results hold within the context of private goods exchange economies (cf., Shapley [55], Shapley and Shubik [60], Champsaur [17], Mas-Colell [43], Cheng [18] and others). Some of these results are for economies without money but all treat private goods exchange economies with divisible goods and concave, monotone utility functions. Moreover, they all treat either replicated sequences of economies or convergent sequences of economies. That games satisfying SGE are asymptotically equivalent to balanced market games clarifies the contribution of the above result. In the context of the prior results developed in this paper, the major shortcoming of the Theorem is that it requires BMC. This author conjectures that the above result, or a close analogue, could be obtained with the milder condition of SGE, but this has not been demonstrated.

Economies with Clubs

By a club economy we mean an economy where participants in the economy form groups – called clubs – for the purposes of collective consumption and/or production collectively with the group members. The groups may possibly overlap. A club structure of the participants in the economy is a covering of the set of players by clubs. Providing utility functions are quasi-linear, such an economy generates a game of the sort discussed in this essay. The worth of a group of players is the maximum total worth that the group can achieve by forming clubs. The most general model of clubs in the literature at this point is Allouch and Wooders [1]. Yet, if one were to assume that utility functions were all quasi-linear and the set of possible types of participants were finite. the results of this paper would apply.

In the simplest case, the utility of an individual depends on the club profile (the numbers of participants of each type) in his club. The total worth of a group of players is the maximum that it can achieve by splitting into clubs. The results presented in this section immediately apply. When there are many participants, club economies can be represented as markets and the competitive payoff vectors for the market are approximated by equal-treatment payoff vectors in approximate cores. Approximate cores converge to equal treatment and competitive equilibrium payoffs. A more general model making these points is treated in Shubik and Wooders [65]. For recent reviews of the literature, see Conley and Smith [19] and Kovalenkov and Wooders [38].[3]

Coalition production economies may also be viewed as club economies. We refer the reader to Böhm [12], Sondermann [73], Shubik and Wooders [70], and for a more recent treatment and further references, Sun, Trockel and Yang [74]).

Let us conclude this section with some historical notes. Club economies came to the attention of the economics profession with the publication of Buchanan [14]. The author pointed out that people care about the numbers of other people with whom they share facilities such as swimming pool clubs. Thus, there may be congestion, leading people to form multiple clubs. Interestingly, much of the recent literature on club economies with many participants and their competitive properties has roots in an older paper, Tiebout [77]. Tiebout conjectured that if public goods are 'local' – that is, subject to exclusion and possibly congestion – then large economies are 'market-like'. A first paper treating club economies with many participants was Pauly [51], who showed that, when all players have the same preferred club size, then the core of economy is nonempty if and only if all participants in the economy can be partitioned into groups of the preferred size. Wooders [82] modeled a club economy as one with local public goods and demonstrated that, when individuals within a club (jurisdiction) are required to pay the same share of the costs of public good provision, then outcomes in the core permit heterogeneous clubs if and only if all types of participants in the same club have the same demands for local public goods and for congestion. Since

[3] Other approaches to economies with clubs/local public goods include Casella and Feinstein [15], Demange [23], Haimanko, O., M. Le Breton and S. Weber [28], and Konishi, Le Breton and Weber [37]. Recent research has treated clubs as networks.

these early results, the literature on clubs has grown substantially.

With a Continuum of Players

Since Aumann [4] much work has been done on economies with a continuum of players. It is natural to question whether the asymptotic equivalence of markets and games reported in this article holds in a continuum setting. Some such results have been obtained.

First, let $N = [01]$ be the 0,1 interval with Lesbegue measure and suppose there is a partition of N into a finite set of subsets $N_1, \ldots, N_T$ where, in interpretation, a point in N_t represents a player of type t. Let Ψ be given. Observe that Ψ determines a payoff for any finite group of players, depending on the numbers of players of each type. If we can aggregate partitions of the total player set into finite coalitions then we have defined a game with a continuum of players and finite coalitions.

For a partition of the continuum into finite groups to 'make sense' economically, it must preserve the relative scarcities given by the measure. This was done in Kaneko and Wooders [35]. To illustrate their idea of measurement consistent partitions of the continuum into finite groups, think of a census form that requires each three-person household to label the players in the household, #1, #2, or #3. When checking the consistency of its figures, the census taker would expect the numbers of people labeled #1 in three-person households to equal the numbers labeled #2 and #3. For consistency, the census taker may also check that the number of first persons in three-person households in a particular region is equal to the number of second persons and third persons in three person households in that region. It is simple arithmetic. This consistency should also hold for k-person households for any k. Measurement consistency is the same idea with the work "number" replaced by "proportion" or "measure".

One can immediately apply results reported above to the special case of TU games of Kaneko–Wooders [35] and conclude that games satisfying small group effectiveness and with a continuum of players have nonempty cores and that the payoff function for the game is one-homogeneous. (We note that there have been a number of papers investigating cores of games with a continuum of players that have came to the conclusion that non-emptiness of exact cores does not hold, even with balancedness assumptions, cf., Weber [78,79]). The results of Wooders [91], show that the continuum economy must be representable by one where all players have the same concave, continuous one-homogeneous utility functions. Market games with a continuum of players and a finite set of types are also investigated in Azriel and Lehrer [3], who confirm these conclusions.)

Other Related Concepts and Results

In an unpublished 1972 paper due to Edward Zajac [97], which has motivated a large amount of literature on 'subsidy-free pricing', cost sharing, and related concepts, the author writes:

> "A fundamental idea of equity in pricing is that 'no consumer group should pay higher prices than it would pay by itself...'. If a particular group is paying a higher price than it would pay if it were severed from the total consumer population, the group feels that it is subsidizing the total population and demands a price reduction".

The "dual" of the cost allocation problem is the problem of surplus sharing and subsidy-free pricing.[4] Tauman [75] provides a excellent survey. Some recent works treating cost allocation and subsidy free-pricing include Moulin [47,48]. See also the recent notion of "Walras' core" in Qin, Shapley and Shimomura [52].

Another related area of research has been into whether games with many players satisfy some notion of the Law of Demand of consumer theory (or the Law of Supply of producer theory). Since games with many players resemble market games, which have the property that an increase in the endowment of a commodity leads to a decrease in its price, such a result should be expected. Indeed, for games with many players, a Law of Scarcity holds – if the numbers of players of a particular type is increased, then core payoffs to players of that type do not increase and may decrease. (This result was observed by Scotchmer and Wooders [54]). See Kovalenkov and Wooders [38,41] for the most recent version of such results and a discussion of the literature. Laws of scarcity in economies with clubs are examined in Cartwright, Conley and Wooders [16].

Some Remarks on Markets and More General Classes of Economies

Forms of the equivalence of outcomes of economies where individuals have concave utility functions but not necessarily linear in money. These include Billera [10], Billera and Bixby [11] and Mas-Colell [42]. A natural question is whether the results reported in this paper can extend to nontransferable utility games and economies where individuals have utility functions that are not necessarily liner

[4] See, for example Moulin [47,48] for excellent discussions of these two problems.

in money. So far the results obtained are not entirely satisfactory. Nonemptiness of approximate cores of games with many players, however, holds in substantial generality; see Kovalenkov and Wooders [40] and Wooders [95].

Conclusions and Future Directions

The results of Shapley and Shubik [60], showing equivalence of structures, rather than equivalence of outcomes of solution concepts in a fixed structure (as in [4], for example) are remarkable. So far, this line of research has been relatively little explored. The results for games with many players have also not been fully explored, except for in the context of games, such as those derived from economies with clubs, and with utility functions that are linear in money.

Per capita boundedness seems to be about the mildest condition that one can impose on an economic structure and still have scarcity of per capita resources in economies with many participants. In economies with quasi-linear utilities (and here, I mean economies in a general sense, as in the glossary) satisfying per capita boundedness and where there are many substitutes for each type of participant, then as the number of participants grows, these economies resemble or (as if they) *are* market economies where individuals have continuous, and monotonic increasing utility functions. Large groups cannot influence outcomes away from outcomes in the core (and outcomes of free competition) since large groups are not significantly more effective than many small groups (from the equivalence, when each player has many close substitutes, between per capita boundedness and small group effectiveness).

But if there are not many substitutes for each participant, then, as we have seen, per capita boundedness allows small groups of participants to have large effects and free competition need not prevail (cores may be empty and price-taking equilibrium may not exist). The condition required to ensure free competition in economies with many participants, without assumptions of "thickness", is precisely small group effectiveness.

But the most complete results relating markets and games, outlined in this paper, deal with economies in which all participants have utility functions that are linear in money and in games with side payments, where the worth of a group can be divided in any way among the members of the group without any loss of total utility or worth. Nonemptiness of approximate cores of large games without side payments has been demonstrated; see Wooders [83,95] and Kovalenkov and Wooders [40]. Moreover, it has been shown that when side payments are 'limited' then approximate cores of games without side payments treat similar players similarly [39].

Results for *specific economic structures*, relating cores to price taking equilibrium treat can treat situations that are, in some respects, more general. A substantial body of literature shows that certain classes of club economies have nonempty cores and also investigates price-taking equilibrium in these situations. Fundamental results are provided by Gale and Shapley [25], Shapley and Shubik [61], and Crawford and Kelso [21] and many more recent papers. We refer the reader to Roth and Sotomayor [53] and to ► Two-Sided Matching Models, by Ömer and Sotomayor in this encyclopedia. A special feature of the models of these papers is that there are two sorts of players or two sides to the market; examples are (1) men and women, (2) workers and firms, (3) interns and hospitals and so on.

Going beyond two-sided markets to clubs in general, however, one observes that the positive results on nonemptiness of cores and existence of price-taking equilibria only holds under restrictive conditions. A number of recent contributions however, provide specific economic models for which, when there are many participants in the economy, as in exchange economies it holds that price-taking equilibrium exists, cores are non-empty, and the set of outcomes of price-taking equilibrium are equivalent to the core. (see, for example, [1,2,24,85,92]).

Bibliography

1. Allouch N, Wooders M (2008) Price taking equilibrium in economies with multiple memberships in clubs and unbounded club sizes. J Econ Theor 140:246–278
2. Allouch N, Conley JP, Wooders M (2008) Anonymous price taking equilibrium in Tiebout economies with a continuum of agents: Existence and characterization. J Math Econ. doi:10.1016/j.jmateco.2008.06.003
3. Azrieli Y, Lehrer E (2007) Market games in large economies with a finite number of types. Econ Theor 31:327–342
4. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
5. Aumann RJ (1987) Game theory. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave: A Dictionary of Economics. Palgrave Macmillan, Basingstoke
6. Aumann RJ, Dreze J (1974) Cooperative games with coalition structures. Int J Game Theory 3:217–37
7. Aumann RJ, Shapley S (1974) Values of Non-Atomic Games. Princeton University Press, Princeton
8. Bennett E, Wooders M (1979) Income distribution and firm formation. J Comp Econ 3:304–317. http://www.myrnawooders.com/
9. Bergstrom T, Varian HR (1985) When do market games have transferable utility? J Econ Theor 35(2):222–233
10. Billera LJ (1974) On games without side payments arising from a general class of markets. J Math Econ 1(2):129–139

11. Billera LJ, Bixby RE (1974) Market representations of n-person games. Bull Am Math Soc 80(3):522–526
12. Böhm V (1974) The core of an economy with production. Rev Econ Stud 41:429–436
13. Bondareva O (1963) Some applications of linear programming to the theory of cooperative games. Problemy kibernetiki 10 (in Russian, see English translation in Selected Russian papers in game theory 1959–1965, Princteon University Press, Princeton
14. Buchanan J (1965) An economic theory of clubs. Economica 33:1–14
15. Casella A, Feinstein JS (2002) Public goods in trade on the formation of markets and jurisdictions. Intern Econ Rev 43:437–462
16. Cartwright E, Conley J, Wooders M (2006) The Law of Demand in Tiebout Economies. In: Fischel WA (ed) The Tiebout Model at 50: Essays in Public Economics in honor of Wallace Oates. Lincoln Institute of Land Policy, Cambridge
17. Champsaur P (1975) Competition vs. cooperation. J Econ Theory 11:394–417
18. Cheng HC (1981) On dual regularity and value convergence theorems. J Math Econ 8:37–57
19. Conley J, Smith S (2005) Coalitions and clubs; Tiebout equilibrium in large economies. In: Demange G, Wooders M (eds) Group Formation in Economies; Networks, Clubs and Coalitions. Cambridge University Press, Cambridge
20. Conley JP, Wooders M (1995) Hedonic independence and taste-homogeneity of optimal jurisdictions in a Tiebout economy with crowding types. Ann D'Econ Stat 75/76:198–219
21. Crawford VP, Kelso AS (1982) Job matching, coalition formation, and gross substitutes. Econornetrica 50:1483–1504
22. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
23. Demange G (1994) Intermediate preferences and stable coalition structures. J Math Econ 1994:45–48
24. Ellickson B, Grodal B, Scotchmer S, Zame W (1999) Clubs and the market. Econometrica 67:1185–1218
25. Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
26. Garratt R, Qin C-Z (1997) On a market for coalitions with indivisible agents and lotteries, J Econ Theor 77(1):81–101
27. Gillies DB (1953) Some theorems on *n*-person games. Ph.D Dissertation, Department of Mathematics. Princeton University, Princeton
28. Haimanko O, Le Breton M, Weber S (2004) Voluntary formation of communities for the provision of public projects. J Econ Theor 115:1–34
29. Hildenbrand W (1974) Core and Equilibria of a Large Economy. Princeton University Press, Princeton
30. Hurwicz L, Uzawa H (1977) Convexity of asymptotic average production possibility sets. In: Arrow KJ, Hurwicz L (eds) Studies in Resource Allocation Processes. Cambridge University Press, Cambridge
31. Kalai E, Zemel E (1982) Totally balanced games and games of flow. Math Oper Res 7:476–478
32. Kalai E, Zemel E (1982) Generalized network problems yielding totally balanced games. Oper Res 30:998–1008
33. Kaneko M, Wooders M (1982) Cores of partitioning games. Math Soc Sci 3:313–327
34. Kaneko M, Wooders M (2004) Utility theories in cooperative games. In: Handbook of Utility Theory vol 2, Chapter 19. Kluwer Academic Press, Dordrecht, pp 1065–1098
35. Kaneko M, Wooders M (1986) The core of a game with a continuum of players and finite coalitions; the model and some results. Math Soc Sci 12:105–137. http://www.myrnawooders.com/
36. Kannai Y (1972) Continuity properties of the core of a market. Econometrica 38:791–815
37. Konishi H, Le Breton M, Weber S (1998) Equilibrium in a finite local public goods economy. J Econ Theory 79:224–244
38. Kovalenkov A, Wooders M (2005) A law of scarcity for games. Econ Theor 26:383–396
39. Kovalenkov A, Wooders M (2001) Epsilon cores of games with limited side payments: nonemptiness and equal treatment. Games Econ Behav 36(2):193–218
40. Kovalenkov A, Wooders M (2003) Approximate cores of games and economies with clubs. J Econ Theory 110:87–120
41. Kovalenkov A, Wooders M (2006) Comparative statics and laws of scarcity for games. In Aliprantis CD, Matzkin RL, McFadden DL, Moore JC, Yannelis NC (eds) Rationality and Equilibrium: A Symposium in Honour of Marcel K. Richter. Studies in Economic Theory Series 26. Springer, Berlin, pp 141–169
42. Mas-Colell A (1975) A further result on the representation of games by markets. J Econ Theor 10(1):117–122
43. Mas-Colell A (1977) Indivisible commodities and general equilibrium theory. J Econ Theory 16(2):443–456
44. Mas-Colell A (1979) Competitive and value allocations of large exchange economies. J Econ Theor 14:307–310
45. Mas-Colell A (1980) Efficiency and decentralization in the pure theory of public goods. Q J Econ 94:625–641
46. Mas-Colell A (1985) The Theory of General Economic Equilibrium. Economic Society Publication No. 9. Cambridge University Press, Cambridge
47. Moulin M (1988) Axioms of Cooperative Decision Making. Econometric Society Monograph No. 15. Cambridge Press, Cambridge
48. Moulin H (1992) Axiomatic cost and surplus sharing. In: Arrow K, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare, 1st edn, vol 1, chap 6. Elsevier, Amsterdam, pp 289–357
49. von Neumann J, Morgenstern O (1953) Theory of Games and Economic Behavior. Princeton University Press, Princeton
50. Owen G (1975) On the core of linear production games. Math Program 9:358–370
51. Pauly M (1970) Cores and clubs. Public Choice 9:53–65
52. Qin C-Z, Shapley LS, Shimomura K-I (2006) The Walras core of an economy and its limit theorem. J Math Econ 42(2):180–197
53. Roth A, Sotomayer M (1990) Two-sided Matching; A Study in Game-theoretic Modeling and Analysis. Cambridge University Press, Cambridge
54. Scotchmer S, Wooders M (1988) Monotonicity in games that exhaust gains to scale. IMSSS Technical Report No. 525, Stanford University
55. Shapley LS (1964) Values of large games -VII: A general exchange economy with money. Rand Memorandum RM-4248-PR
56. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 9:45–48
57. Shapley LS (1952) Notes on the N-Person game III: Some variants of the von-Neumann-Morgenstern definition of solution. Rand Corporation research memorandum, RM-817:1952

58. Shapley LS, Shubik M (1960) On the core of an economic system with externalities. Am Econ Rev 59:678–684
59. Shapley LS, Shubik M (1966) Quasi-cores in a monetary economy with nonconvex preferences. Econometrica 34:805–827
60. Shapley LS, Shubik M (1969) On market games. J Econ Theor 1:9–25
61. Shapley LS, Shubik M (1972) The Assignment Game 1; The core. Int J Game Theor 1:11–30
62. Shapley LS, Shubik M (1975) Competitive outcomes in the cores of market games. Int J Game Theor 4:229–237
63. Shapley LS, Shubik M (1977) Trade using one commodity as a means of payment. J Political Econ 85:937–68
64. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40. Princeton University Press, Princeton, pp 267–278
65. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
66. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
67. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40, Princeton University Press, Princeton, pp 267–278
68. Shubik M, Wooders M (1982) Near markets and market games. Cowles Foundation Discussion Paper No. 657, on line at http://www.myrnawooders.com/
69. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
70. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part I Replica games, externalities, and approximate cores. Math Soc Sci 6:27–48
71. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
72. Shubik M, Wooders M (1986) Near-markets and market-games. Econ Stud Q 37:289–299
73. Sondermann D (1974) Economics of scale and equilibria in coalition production economies. J Econ Theor 8:259–291
74. Sun N, Trockel W, Yang Z (2008) Competitive outcomes and endogenous coalition formation in an n-person game. J Math Econ 44:853–860
75. Tauman Y (1987) The Aumann–Shapley prices: A survey. In: Roth A (ed) The Shapley Value: Essays in Honor of Lloyd S Shapley. Cambridge University, Cambridge
76. Tauman Y, Urbano A, Watanabe J (1997) A model of multiproduct price competition. J Econ Theor 77:377–401
77. Tiebout C (1956) A pure theory of local expenditures. J Political Econ 64:416–424
78. Weber S (1979) On ε-cores of balanced games. Int J Game Theor 8:241–250
79. Weber S (1981) Some results on the weak core of a non-sidepayment game with infinitely many players. J Math Econ 8:101–111
80. Winter E, Wooders M (1990) On large games with bounded essential coalition sizes. University of Bonn Sondeforschungsbereich 303 Discussion Paper B-149. on-line at http://www.myrnawooders.com/ Intern J Econ Theor (2008) 4:191–206
81. Wooders M (1977) Properties of quasi-cores and quasi-equilibria in coalition economies. SUNY-Stony Brook Department of Economics Working Paper No. 184, revised (1979) as A characterization of approximate equilibria and cores in a class of coalition economies. State University of New York Stony Brook Economics Department. http://www.myrnawooders.com/
82. Wooders M (1978) Equilibria, the core, and jurisdiction structures in economies with a local public good. J Econ Theor 18:328–348
83. Wooders M (1983) The epsilon core of a large replica game. J Math Econ 11:277–300, on-line at http://www.myrnawooders.com/
84. Wooders M (1988) Large games are market games 1. Large finite games. C.O.R.E. Discussion Paper No. 8842 http://www.myrnawooders.com/
85. Wooders M (1989) A Tiebout Theorem. Math Soc Sci 18: 33–55
86. Wooders M (1991) On large games and competitive markets 1: Theory. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. (B-195, Revised August 1992). http://www.myrnawooders.com/
87. Wooders M (1991) The efficaciousness of small groups and the approximate core property in games without side payments. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. B-179. http://www.myrnawooders.com/
88. Wooders M (1992) Inessentiality of large groups and the approximate core property; An equivalence theorem. Econ Theor 2:129–147
89. Wooders M (1992) Large games and economies with effective small groups. University of Bonn Sonderforschingsbereich 303 Discussion Paper No. B-215.(revised) in Game-Theoretic Methods in General Equilibrum Analysis. (eds) Mertens J-F, Sorin S, Kluwer, Dordrecht. http://www.myrnawooders.com/
90. Wooders M (1993) The attribute core, core convergence, and small group effectiveness; The effects of property rights assignments on the attribute core. University of Toronto Working Paper No. 9304
91. Wooders M (1994) Equivalence of games and markets. Econometrica 62:1141–1160. http://www.myrnawooders.com/n
92. Wooders M (1997) Equivalence of Lindahl equilibria with participation prices and the core. Econ Theor 9:113–127
93. Wooders M (2007) Core convergence in market games and club economics. Rev Econ Design (to appear)
94. Wooders M (2008) Small group effectiveness, per capita boundedness and nonemptiness of approximate cores. J Math Econ 44:888–906
95. Wooders M (2008) Games with many players and abstract economies permitting differentiated commodities, clubs, and public goods (submitted)
96. Wooders M, Zame WR (1987) Large games; Fair and stable outcomes. J Econ Theor 42:59–93
97. Zajac E (1972) Some preliminary thoughts on subsidization. presented at the Conference on Telecommunications Research, Washington

Market Microstructure

Clara Vega, Christian S. Miller[1]
Division of International Finance,
Board of Governors of the Federal Reserve System,
Washington DC, USA

Article Outline

Glossary

Ask price Price at which a trader is willing to sell an asset. The most competitive ask price in a financial market or best ask is the lowest price offered by a seller.

Bid price Price at which a trader is willing to buy an asset. The most competitive bid price or best bid in a financial market is the highest price offered by a buyer.

Limit order Orders placed by market participants contingent upon the realization of a certain price in the market. In other words, traders will identify a maximum or minimum price at which they are willing to buy or sell a specific quantity of a particular asset.

Market order Order to buy or sell a particular asset immediately at current market prices.

Market structure The way in which trade occurs within a particular market. Institutions have constructed idiosyncratic guidelines to dictate how transactions can take place, so generalizing one trading structure to model all markets is quite difficult, if impossible.

Order flow is the cumulative flow of signed transactions over a time period, where each transaction is signed positively or negatively depending on whether the initiator of the transaction (the non-quoting counterparty) is buying or selling, respectively. By definition, in any market, the quantity purchased of an asset equals the quantity sold of the same asset. The key is to sign the transaction volume from the perspective of the initiator of the transaction.

Bid-ask spread The difference between the highest bid price and the lowest ask price. This difference, or spread, constitutes part of the cost of trading.

Definition of the Subject

Market microstructure is a field of study in economics that examines the way in which assets are traded and priced under different trading mechanisms, e. g., single-price call auction, dealer markets, limit-order book markets, hybrid markets, etc., and under different trading environments, e. g., perfect information environments (complete markets) compared to asymmetric information environments (incomplete markets). While much of economics abstracts from the market structure and market frictions the microstructure literature specializes in understanding them and the effects they may have on asset prices and quantities traded. Even though economic theorists assume a frictionless economy to prove powerful theorems about the efficiency of a decentralized market system, the market structure and market frictions can be very important. Ignoring them may lead researchers and policy makers to wrong conclusions. For example, in a Walrasian world with perfect information and no transaction costs, prices efficiently aggregate information when trading is organized as a single-price call auction with large numbers of traders. However, most securities markets are not single-price call auctions as several studies show that this trading mechanism may be optimal when uncertainty about the fundamental value of the asset is high, but it is not optimal at other times. Furthermore, in the 1970's the economics of information literature argued that allowing for imperfect information could overturn the central implication of the complete-markets model, that competitive, decentralized markets yield economically efficient results.

Market Structures

A large part of the market microstructure field consists of developing models to describe the behavior of individuals acting according to the guidelines of various trading institutions, and to study how trading quantities and prices in various markets arise given a particular set of assumptions. Thus, we start with a short description of the common market structures. It is outside the scope of this article to detail the myriad rules that govern various financial markets. It is also counter-productive because trading systems are in a continuous process of structural changes generated by research, competition, and technological innova-

[1] The concepts in this paper in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

tions. Instead we present a general outline of the guidelines that dictate the way in which assets trade and the effects these rules may have on asset prices and quantities traded.

Auctions

Auctions are order-driven trading mechanism, i. e., investors submit their orders before observing the transaction price. In contrast, investors in a quote-driven trading mechanism obtain firm price quotations from dealers prior to order submission (these price quotations usually depend on the size of the order). Auctions can be continuous or periodic. An example of a continuous auction is the automated limit order book, which consists of a sequence of bilateral transactions at possibly different prices (we describe limit-order books in more detail below). In contrast, a periodic or call auction is characterized by multilateral transactions. Periodic or batch systems, such as the single-price call auction, are used to set opening prices in several exchanges, e. g., NYSE, Tokyo Stock Exchange, etc. In these markets limit orders and market-on-open orders are collected overnight. At the beginning of the trading day the specialist chooses the price that enables the largest number of orders to be executed. Stock exchanges use call auctions to fix opening prices because uncertainty about fundamentals is larger at the opening than during regular trading hours. Indeed, Madhavan [36] provides a theoretical argument for batch markets as a way to reduce market failures caused by information asymmetries. Another example of a periodic auction market is the primary market for US Treasury securities. These securities are sold through sealed-bid single-price auctions at pre-determined dates announced by the US Treasury Department (before November 1998 the Treasury auctioned securities through multiple-price or discriminatory auctions).

Limit Order Markets

Limit order books are the most widespread conduit to facilitate trade in financial markets; at least one limit order book exists in most continuous (liquid) security markets (see p. 10 in [29]). In such markets, traders submit their bid and ask orders, and the order book(s) process these orders, comparing them to already existing orders to establish whether any matches can be made. These pre-existing, unfilled limit orders comprise the limit order book. Various rules dictate how and when limit orders are acted upon Parlour and Seppi [42]. Generally price and then time determine priority of execution. For instance, a limit order to sell an asset for $50 will take precedence over an order to sell at $52. If two limit orders are priced the same, then the first limit order submitted is the first order executed.

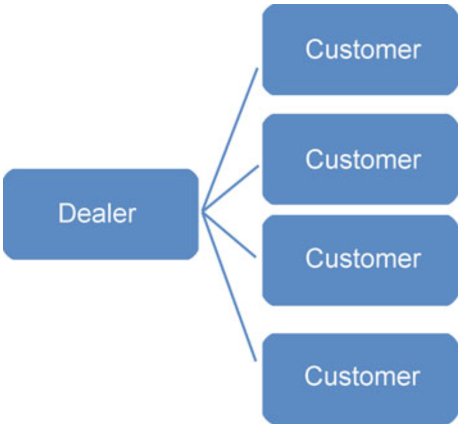
Sometimes traders may request the execution of a market order; this order is immediately executed at the best price available. A problem can arise if the quantity designated in the market order is larger than the quantity available at the best price available on the limit book. Different exchanges have different rules to deal with the left-over quantity. In the NYSE, the excess quantity "walks the book", meaning that the market order achieves partial executions at progressively worse prices until the order is filled. This process results in the partial execution of market orders at less than desirable prices for the order-issuing trader. In contrast, in the Euronext system and the Paris Bourse, if the quantity in the market order exceeds the quantity at the best price, the unfilled part of the market order is transformed into a limit order, requesting execution on the remaining quantity at the same execution price.

Various other rules regarding the execution of limit orders exist. For example, traders can post orders with an expiration time, i. e., the limit order is canceled if it is not executed within a given time frame. This prevents limit orders to be "picked off" by investors who receive updated public or private information. Traders can also hide part of the order they submit to the limit order book, these are called "iceberg" orders.

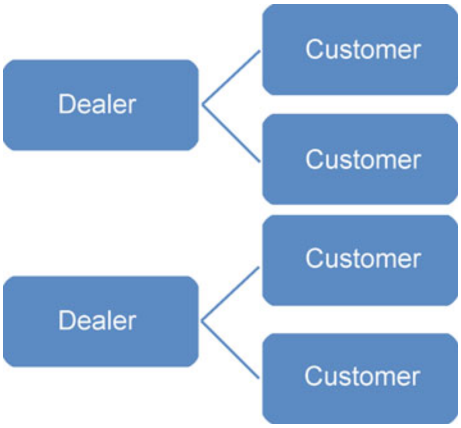
Exchanges vary in the degree of transparency of the limit order books. The automated limit-order-book system used by the Toronto Stock Exchange and the Paris Bourse are among the most transparent systems. They offer continuous trading and the public display of current and away limit orders (an open book limit-order system). The NYSE has shifted from a close limit order book policy (although specialists made the book available to traders on the floor at their own discretion) to making the content of the limit-order book public. In January 24, 2002 the service OpenBook was introduced. This service provides information about depth in the book in real time at each price level for all securities to subscribers either directly from the NYSE or through data vendors such as Reuters and Bloomberg. Boehmer et al. [10] empirically examine the effect of increased transparency in the NYSE and Goettler et al. [24] numerically solves a dynamic model of limit orders in which agents arrive randomly and may trade one share in an open electronic limit order market.

Single Dealer Markets

It is a market where one dealer (market maker or specialist) stands ready to buy at his bid quote and sell at his offer quote. In this environment, incoming orders are necessarily market orders (in contrast to limit orders). The



Market Microstructure, Figure 1
Single dealer market



Market Microstructure, Figure 2
Multiple dealer market

customer either buys (sells) at the dealer's offer (bid) or chooses not to trade. Dealer markets are less transparent than open book limit order markets (only the best-bid and best-ask price are known to the customer in a dealer market, while the entire depth of the market is visible in an open limit-order book). In reality there are very few pure single-dealer markets. The NYSE is sometimes mistakenly labeled as a single-dealer market, but it is a hybrid system with both limit-order and single-dealer features. Equity trading is centered on the stock specialist, who is assigned particular stocks in which to make a market. Each listed security has a single specialist, and all trading on the exchange must go through the specialist. The specialist receives market orders (orders for immediate execution) and limit orders (orders contingent on price, quantity and time), so that specialists do not enjoy monopoly power because they compete against the limit order book. If a market order comes to buy, the specialist can either match it with the best sell limit order or if he offers a lower price, he can take the other side. Examples of pure single-dealer markets are foreign exchange markets in developing countries with fixed exchange rates, where all orders must be routed through a single dealer – the central bank.

Multiple Dealer Markets

Competition in this environment is brought through multiple dealers. In a centralized market, quotes from many dealers are available on a screen (NASDAQ) or on the floor of an exchange (like a futures trading pit: the Chicago Board of Trade, the New York Mercantile Exchange, and the Chicago Mercantile Exchange). In a decentralized market trading occurs over-the-counter rather than through an organized exchange. The foreign exchange market, government bond's secondary market and corporate bond markets are good examples of decentralized multiple-dealer markets. There is less transparency in these markets than in a centralized multiple-dealer one because not all dealer quotes are observable. As a result, there can be simultaneous transactions that occur at different prices. The main mechanism that mitigates the dealer's monopoly power is the fact that the interaction between a dealer and a customer is repeated over time. Dealers have an incentive to keep their reputation in quoting reasonable bid and ask prices so that the customer does not go to another dealer. In particular, dealers are concerned about losing large customers, so that small customers have less bargaining power. Competition in these markets and pressure from regulators has also forced a shift from voice-based brokers to electronic brokers, who provide a higher level of transparency. For example, recently the Bond Market Association responded to SEC pressure for more transparency in the corporate bond market by setting up a single reporting system for investment grade bonds Viswanathan and Wang [47]. For a detailed description of how the foreign exchange market and the government bond market operate please refer to Lyons [35] and Fabozzi and Fleming [21], respectively.

Inter-Dealer Markets

In addition to dealer-customer interactions, inter-dealer trading is very important. Ho and Stoll [31] suggest that risk-sharing is the main reason for inter-dealer trading. The incoming orders that a particular dealer receives are rarely balanced, so that the dealer is left with an undesired short or long position. To balance their inventory the dealer can sell to or buy from other dealers. The dealer can do so by either contacting another dealer directly or through a broker. The benefit of going through a broker is that they provide anonymity and the cost is the broker fee. In addition, brokers offer dealers electronic trading platforms that help the flow of information. These screens typically post: (i) the best bid and offer prices of several dealers, (ii) the associated quantities bid or offered, (iii) transaction prices, and (iv) transaction size. Common brokers in the secondary government bond market are ICAP's BrokerTec, Cantor Fitzgerald/eSpeed, Garban-Intercapital, Hilliard Farber, and Tullett Liberty. The main electronic brokers in the major spot markets (JPY/USD, Euro/USD, CHF/USD and GBP/USD currency pairs) are EBS and Dealing 2000-2, a dealer-broker Reuter product (Dealing 2000-1 is the Reuter product for direct inter-dealer trading). It is worth noting that EBS and Dealing 2000-2 typically conduct trades via a limit order book, while Reuters D2000-1 is a sequential trading system (an outside customer trades with dealer 1 who trades with dealer 2 who trades with dealer 3 and so on; hence it is often referred to as "hot potato" trading). In the equity markets inter-dealer trading is also common. On the NASDAQ market, dealers can trade with each other on the SuperSoes system, the SelectNet system and on electronic crossing networks (ECNs) like Instinet. In equity markets like the London Stock Exchange, inter-dealer trading constitutes about 40% of the total volume Viswanathan and Wang [47], while in the foreign exchange market and the US government bond market inter-dealer trading far exceeds public trades. Inter-dealer trading accounts for about 85% Lyons [35] of the trading volume in the foreign exchange market and about 99% Viswanathan and Wang [47] in the US government bond market. Two-thirds of the transactions in the US government bond market are handled by inter-dealer brokerage firms and the remaining one-third is done via direct interactions between the primary dealers listed in Table 1. For more details on inter-dealer trading please refer to Viswanathan and Wang [47].

In the next section we present a few of the basic models that are employed in the market microstructure literature.

Market Microstructure, Table 1

Primary Government Securities Dealers as of Nov. 30, 2007. Source: Federal Reserve Bank of New York
http://www.newyorkfed.org/markets/pridealers_listing.html

BNP Paribas Securities Corp.
Banc of America Securities LLC
Barclays Capital Inc.
Bear, Stearns & Co., Inc.
Cantor Fitzgerald & Co.
Citigroup Global Markets Inc.
Countrywide Securities Corporation
Credit Suisse Securities (USA) LLC
Daiwa Securities America Inc.
Deutsche Bank Securities Inc.
Dresdner Kleinwort Wasserstein Securities LLC.
Goldman, Sachs & Co.
Greenwich Capital Markets, Inc.
HSBC Securities (USA) Inc.
J.P. Morgan Securities Inc.
Lehman Brothers Inc.
Merrill Lynch Government Securities Inc.
Mizuho Securities USA Inc.
Morgan Stanley & Co. Incorporated
UBS Securities LLC.

Inventory Models

The first theoretical models in the market microstructure field were inventory models; however information-based models have come to dominate the field because the former describe temporary price deviations around the equilibrium price, while the later describe permanent price changes. The main idea of inventory models is captured by Smidt [44] who argued that dealers, or market makers in general, are not simply passive providers of immediacy, but actively adjust the bid-ask spread in response to fluctuation in their inventory levels. Though dealers' main responsibility is to facilitate trade in an asset market, they set prices to realize rapid inventory turnover and to prevent the accumulation of significant positions on one side of the market. The consequence of this paradigm is a price that may diverge from the expected value of an asset if a dealer is long or short relative to a desired (target) inventory, which would result in temporary price movements over various (short-term) periods of time. How "short-term" these deviations are differs across studies. Data on specialists' inventories is scarce, but studies have been successful in showing that inventories play an important role in intraday trading and a recently published paper by Hendershot and Seasholes [30] shows that inventory considerations affect prices beyond intraday trading. Hendershott and Seasholes argument is that market makers are willing to provide liquidity as long as they are able to buy (sell) at a discount (premium) relative to future prices. Hence, large inventories of the market maker should coincide with large buying or selling pressure, which cause prices to subsequently reverse (e. g., Amihud and Mendelson [2] and Grossman and Miller [26] provide inventory models that lead to reversals). But the reversal of prices does not have to be immediate, in fact, they document that reversals can take as long as 12-days.

Inventory models assume that there is no asymmetric information. Fluctuations in market prices, therefore, results solely from dealers' decisions about the positions of their inventory. Dealers' hold sub-optimal portfolios, bare a cost for maintaining inventories – holding assets for the purpose of providing liquidity to the market exposes them to risk. Consequently, market makers receive compensation (i. e., bid-ask spread) for incurring the transaction costs entailed in managing their inventories.

Various texts, including O'Hara [41], present different inventory models. The discussion below will focus on one such model—the model presented by Garman [22] that inaugurated the field of market microstructure and builds on Smidt [44] idea. As O'Hara notes, aspects of basic inventory models, such as the assumption of perfect information, are not realistic; however, it is still useful to review basic models' assumptions about the functioning of asset markets to isolate the various ways in which the behavior of market makers can influence asset prices.

Garman's Model

The expected value of the asset or the equilibrium price is equal to the price at which quantity demanded equals quantity supplied at a particular period in time. Let's label this price p^*. Garman (16) shows that it is optimal for the market maker to charge two different prices. One price, p_a, the ask price, at which he will fill orders wishing to buy the stock, and another price, p_b, the bid price, at which he will fill order wishing to sell the stock. These prices will not necessarily straddle the equilibrium price, p^*, i. e., $p_b > p^* > p_a$. By being willing to take profits in the form of stock inventory increases, the market maker can artificially inflate prices by maintaining the inequality $p_b > p_a > p^*$. In no case, however, will the market maker be able to set both prices below p^* without ultimately running out of inventory. Furthermore, if the market maker sets both prices equal to each other, equal to the equilibrium price, i. e., $p_b = p^* = p_a$, then the market maker will fail with certainty (i. e., the market maker will either run out of inventory or cash with probability equal to 1). In what follows we describe briefly how the model works and we ask the reader to refer to the original paper for more details. Garman considered two market clearing mechanisms: a dealer structure and a double auction mechanism; however, we will focus on the dealer structure only.

Garman conceived the dealer as a monopolist; he alone receives orders from traders, determines asset prices, and facilitates trade. In making the market, the dealer engages in optimizing behavior by maximizing his expected profit per unit of time while avoiding bankruptcy or failure, which is defined as depleting his inventory or losing all of his money. The dealer sets an ask price and a bid price at the beginning of trading, and investors submit their orders after observing the dealer's bid and ask quote. The arrivals of orders to buy and sell the asset are independent stochastic processes that are distributed according to a Poisson distribution. The dealer, therefore, runs a chance of failing since he must ensure liquidity—selling part of his inventory or buying a particular asset as determined by the arrival rate of buyers and sellers.

Assuming a Poisson arrival rate necessitates that (i) many agents are interacting in the market, (ii) these agents issue orders independently without consideration of others' behavior, (iii) no one agent can issue an infinite number of orders in a finite period, and (iv) no subset of agents can dominate order generation, which precludes the existence of private information. It requires that the order flow be stochastic without being informative about future market or price movements.

Garman's model is based on two equations—one that determines the dealer's cash, $I_c(t)$, at time t and one that determines the dealer's inventory of the asset, $I_s(t)$, at time t. At time 0, the dealer holds $I_c(0)$ units of cash and $I_s(0)$ of stock. Inventories at any point in time can be represented as follows:

$$I_c(t) = I_c(0) + p_a N_a(t) - p_b N_b(t) ,$$
$$I_s(t) = I_s(0) + N_b(t) - N_a(t) ,$$

where $N_a(t)$ is the number of executed buy orders at time t, $N_b(t)$ is the number of executed sell orders at time t, p_a is the ask price and p_b is the bid price for a stock. Using these equations, Garman sets forth to determine how a dealer can avoid market failure or bankruptcy (i. e., $I_c(t)$ or $I_s(t) = 0$). Preventing this situation from occurring is the main goal of dealers in setting asset prices. Garman [22] provides a detailed explanation for determining when failure will occur, but for the purpose of this article it is enough to skip to the main conclusion. In order to avoid market failure, dealers must set p_a and p_b to satisfy both equations:

$$p_a\lambda_a(p_a) > p_b\lambda_b(p_b) \quad \text{and}$$
$$\lambda_b(p_b) > \lambda_a(p_a)$$

where $\lambda_a(p_a)$ is the probability of stock leaving the dealer's inventory and $\lambda_b(p_b)$ is the probability of stock being added to the dealers inventory. Simultaneously satisfying these equations requires that the dealer set p_a above p_b.

In other words, a spread must be in place in order for the dealer to avoid bankruptcy or market failure, though the market maker still faces a positive probability of failure.

Various inventory models exist that explain the presence of the bid-ask spread. Although Garman's approach focuses on the threat of market failure to explain the disparity in bid and ask prices, other explanations such as dealers' market power or risk aversion have also been proposed by theorists (see p. 51 in O'Hara [41]). Though the dissimilarities among inventory models are many, the common theme that links these models together is the complex balancing problem faced by the dealer who must moderate random deviations in inflows and outflows of cash and assets. Over the long run the flow of orders had no effect on asset prices, but the dealers' attempt to recalibrate their positions in response to the random stochastic order flows causes price fluctuation in the short run.

Information-Based Models

One implication of the inventory approach discussed in the previous section is that inventory costs determine the bid-ask spread. Beginning with an insightful paper by Bagehot [9], a new theory emerged to explain bid-ask spreads that did not rely on inventory costs, but rather posited an important role for information. These information-based models used insights from the theory of adverse selection to demonstrate how, even in competitive markets without explicit inventory costs, spreads would exist. In what follow we describe three information-based models to illustrate the insights gained from adopting an information-based approach to studying market interactions.

Copeland and Galai's Model

Copeland and Galai [14] were first to construct a formalized model incorporating information costs. Similar to Garman's inventory model the agents in the model are dealers and traders. In contrast to Garman's model, there is more than one dealer and there are two types of traders: informed and uninformed. Informed traders know the true value of the asset, P, and uninformed or liquidity traders trade for exogenous reasons to the value of the asset (e. g., immediate consumption needs). The existence of uninformed traders that trade for non-speculative reasons is ubiquitous in the literature. This assumption is necessary because for information to be valuable informed traders need to be anonymous. If traders known to possess superior knowledge could easily be identified, then no one would agree to trade with them. This is the so called no-trade equilibrium described in Milgrom and Stokey [39].

The trader arrival process is exogeneously determined and is independent of the price change process. This is the same assumption as in Garman's model, but this assumption is not harmless in the presence of informed traders as it appears likely that informed trader behavior would depend on what they know about the true value of the asset relative to what the market thinks. This aspect of the problem is not resolved in Copeland and Galai's paper, but other authors relax this assumption and allow the number of informed traders in the market to be endogenously determined. However, the main contribution of Copeland and Galai's paper is to show that even in the presence of competitive dealers, the mere presence of informed traders implies that the bid-ask spread will be positive. The dealer knows the stochastic process that generates prices, $f(P)$, knows the probability that the next trader is informed, π_1, and knows the elasticity of demand of uninformed and informed traders. With this information the objective of the dealer is to choose a bid-ask spread that maximizes his profits. If the dealer sets the bid-ask spread too wide, he loses expected revenues from uninformed traders, but reduces potential losses to informed traders. On the other hand, if he establishes a spread which is too narrow, the probability of losses incurring to informed traders increases, but is offset by potential revenues from liquidity traders. His optimal bid-ask spread is determinted by a tradeoff between expected gains from liquidity trading and expected losses to informed trading.

The timing of the model is as follows. A trader arrives to the trading post, the dealer offers a quote, and the "true" price, P, is revealed immediately after the trade. An uninformed trader will buy an asset with probability π_{BL}, sell an asset with probability π_{SL}, and decide not to trade with probability π_{NL}. (The "L" in this notation reflects the fact that Copeland and Galai refer to uninformed traders as liquidity traders.) Because informed traders know the true value of P, their decisions to buy, sell, or refrain from trade are based on strategies that maximize their profit.

Dealers at any instant will trade with informed traders with probability π_1 and can expect to lose:

$$\int_{P_{\mathrm{A}}}^{\infty}(P-P_{\mathrm{A}})f(P)\mathrm{d}P+\int_{0}^{P_{\mathrm{B}}}(P_{\mathrm{B}}-P)f(P)\mathrm{d}P\,,$$

where P_{A} and P_{B} are the ask and bid prices quoted by the dealer, and P is the "true" value of the asset. Dealers at any instant will trade with uninformed traders with probability $1-\pi_1$ and can expect to gain:

$$\pi_{\mathrm{BL}}(P_{\mathrm{A}}-P)+\pi_{\mathrm{SL}}(P-P_{\mathrm{B}})+\pi_{\mathrm{NL}}(0)$$

Because the dealer does not know whether individual trades are with informed or uninformed traders, the dealers' objective function is the product of π_1 and the first equation added to the product of $1-\pi_1$ and the second equation. The dealers' optimal bid and ask prices result from this maximization problem. If the prices are negative, however, the market closes.

Not all informed traders who arrive at the marketplace will trade. Informed traders who believe the quoted price by the dealer will fall between P_A and P_B will not trade. Hence, the elasticity of demand by informed traders with respect to the bid-ask spread interval is implicit in the limits of integration in the equation above. The dealers revenue comes from those liquidity traders who are willing to pay $P_\text{A}-P$ or $P-P_\text{B}$ as a price for immediacy. The authors assume that the likelihood that a liquidity trader will consummate trade declines as the bid-ask spread increases, in other words, the liquidity traders elasticity of demand is implicit in the probabilities that liquidity traders will either buy the asset, sell it or not trade.

The framework described above can include competition by incorporating a zero-profit constraint into the dealers problem. The most important result is that even with risk neutral, competitive dealers, the bid-ask spread is positive. The size of the spread will depend on the particular elasticities of the traders' demand functions, and the arrival rate of informed and uninformed traders. As long as there is a positive probability that some trader is informed, the spread will not be zero.

This model, however, is a static one-trade framework and as such it does not allow trade itself to convey information. The model we describe in the next section captures the dynamic aspect of trading and introduces the concept of trade as signals of information.

Easley and O'Hara's Model

What follows is a brief summary of the model; for an extensive discussion of the structure of the model please refer to Easley and O'Hara [16].

The game consists of three players, liquidity traders, informed traders and a market maker. All players are risk neutral, there are no transactions costs, and there is no discounting by traders. The no-discounting assumption is reasonable since agents are optimizing their behavior over one day. Liquidity traders buy or sell shares of the asset for reasons that are exogenous to the model and each buy and sell order arrives to the market according to an independent Poisson distribution with a daily arrival rate equal to ε. The probability that an information event occurs is α, in which case the probability of bad news is δ and the probability of good news is $(1-\delta)$. If an information event occurs, the arrival rate of informed traders is μ. Informed traders trade for speculative reasons; if they receive good news (the current asset price is below the liquidation value of the asset) they buy one share of the asset, if they receive bad news they sell one share of the asset.

On days with no information events, which occur with probability $(1-\alpha)$, the arrival rate of buy orders is ε and the arrival rate of sell orders is ε as well. The model can be parametrized so that the arrival rate of liquidity buyers and sellers is different. However, the numbers of trades for certain stocks from 2000 on are very large, particularly for Nasdaq stocks, and as a result the parameter estimates suffer from a truncation error. To minimize this problem, it is useful to set the arrival rates of liquidity sellers and buyers equal to each other, so that one can factor out a common factor in the likelihood function as in Easley, Engle, O'Hara, and Wu [19]. Figure 3, represents a diagram of how the model works.

Thus, the total amount of transactions on non-information days is 2ε with the number of buys approximately equal to the number of sells. On a bad information event day, which occurs with probability $\alpha\delta$, we observe more sells than buys. To be precise, the arrival rate of buy orders is ε and the arrival rate of sell orders is $\varepsilon+\mu$. In contrast, on a good information event day, which occurs with probability $\alpha(1-\delta)$, we observe more buys than sells, i. e., the arrival rate of buy orders is $\varepsilon+\mu$ and the arrival rate of sell orders is ε.

Easley and O'Hara [16] define PIN as the estimated arrival rate of informed trades divided by the estimated arrival rate of all trades during a pre-specified period of time. Formally,

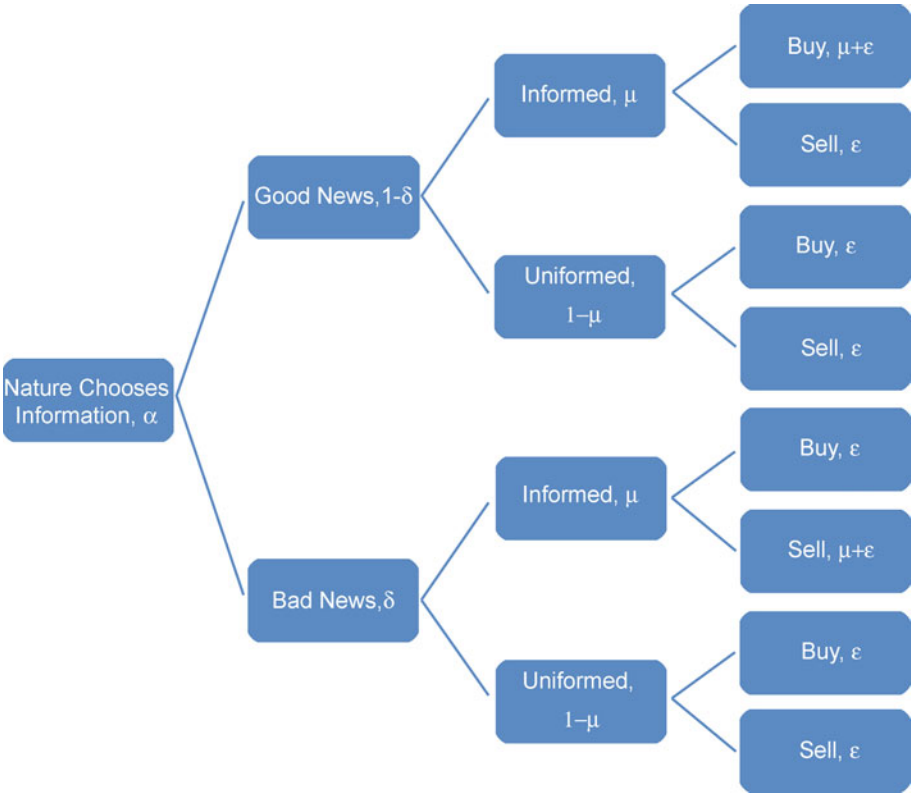
$$\text{PIN} = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu}+2\hat{\varepsilon}}\,.$$

One can estimate all four parameters, $\theta=\{\varepsilon,\mu,\alpha,\delta\}$, by maximizing the likelihood function

$$L(\theta|M)=\prod_{t=1}^{T} L(\theta|B_t,S_t)$$

where B_t is the number of buys and S_t is the number of sells on day t. Assuming days are independent, the likelihood of observing the history of buys and sells $\{M=(B_t,S_t)\}_{t=1}^{T}$ over T days is just the product of the daily likelihoods,

$$\begin{aligned}L(\theta|M) &= \alpha\delta\mathrm{e}^{-(2\varepsilon+\mu)}\frac{\varepsilon^B(\varepsilon+\mu)^S}{B!S!}\\ &\quad+\alpha(1-\delta)\mathrm{e}^{-(2\varepsilon+\mu)}\frac{(\varepsilon+\mu)^B\varepsilon^S}{B!S!}\\ &\quad+(1-\alpha)\delta\mathrm{e}^{-(2\varepsilon)}\frac{\varepsilon^{B+S}}{B!S!}\end{aligned}$$



Market Microstructure, Figure 3
The tree diagram of the trading process [16]

where T is equal to the time frame the researcher is interested in, e. g., Vega [46] choses 40 trading days before an earnings announcement is released, Easley, O'Hara, and Paperman [17] also use 40 trading days to estimate PIN, while Easley, Hvidkjaer, and O'Hara [18] use one calendar year to estimate PIN. The more trading days one uses to estimate PIN the more accurately one will measure information-based trading. Hence, one should check for robustness different estimation windows.

While all the parameters are identified and the likelihood function is differentiable, there is no closed-form solution to the four $(\varepsilon, \mu, \alpha, \delta)$ first-order conditions. Nevertheless, the arrival rate of liquidity traders ε can be interpreted as the daily average number of transactions during the estimation window. The parameter μ reflects the abnormal or unusual number of transactions. The parameter α is equal to the proportion of days characterized by an abnormal level of transactions. The parameter δ is equal to the number of days with an abnormal number of sells divided by the number of days with an abnormal level of transactions.

To calculate the daily number of buys and sells most authors use the Lee and Ready [33] algorithm for NYSE- and AMEX-listed stocks and Ellis, Michaely, and O'Hara's [20] suggested variation of the Lee and Ready algorithm for Nasdaq-listed stocks. Odders–White [40], Lee and Radhakrishna [34], and Ellis, Michaely, and O'Hara [20] evaluate how well the Lee and Ready algorithm performs and they find that the algorithm is from 81% to 93% accurate, depending on the sample period and stocks studied. Thus the measurement error is relatively small.

To estimate the model using US stock market data most researchers use bid quotes, ask quotes, and transaction prices from the Institute for the Study of Securities Markets (ISSM) and the Trade and Quotes (TAQ) database. ISSM data contains tick-by-tick data covering the NYSE and AMEX trades between 1983 to 1992 and NASDAQ trades from 1987 to 1992, while TAQ data covers the sample period from 1993 to the present.

Vega [46] plots the time series of the parameter estimates in addition to the PIN measure averaged across all stocks in the sample. It is evident in that plot that the parameters ε and μ are not stationary. These parameters are related to the trading frequency, hence they are upwards-trending as the number of transactions has increased over the years. In contrast, the estimates of δ, α, and PIN are stationary over the years.

Vega [46] also shows average quarterly bivariate correlations of firm characteristics and PIN. PIN is most highly correlated with log market value with a bivariate correlation coefficient equal to -0.481. The cross-sectional range of -0.70 to -0.32 over the 64 periods implies that across stocks within the same quarter, PIN is negatively correlated with the firm capital size. To test this hypothesis formally Vega [46] first calculate Mann–Whitney test statistics for all periods. Then she tests the hypothesis that the sample of large firms has the same median PIN as the sample of small firms against the alternative hypothesis that they have different medians. In untabulated results she finds that she can soundly reject the null hypothesis in favor of the alternative for 60 out of the 64 periods she analyzes.

The negative relation between private information and firm size is consistent with both previous empirical studies that use PIN as an informed trading measure and Diamond and Verrecchia [15] who assert that asymmetric information is largest for small firms.

Next we present the Kyle Model, which is a workhorse within the market microstructure literature.

Kyle Model

In this information model, an auctioneer determines a price after all traders, uninformed and informed, submit their orders. Besides the risk-neutral market-maker, there is also one risk-neutral informed trader and multiple uninformed traders, who do not issue strategic orders. The market makers are unable to distinguish orders emanating from informed traders from those issued by uninformed traders. Informed traders understand this lack of transparency and can use it for their own advantage.

In the Kyle model there is just one risky asset that is traded over one period. This period of time consists of four distinct phases. First, the informed trader (and only the informed trader) observes the value V of the risky asset's payoff at the end of the period. V is a normally distributed random variable with mean zero and variance equal to σ_v^2. Second, market orders from the informed trader as well as the uninformed traders are submitted to the auctioneer, who is unaware of the end-of-period payoff of the asset, V. The market orders from the informed trader can be represented by D^I, and the market orders from the uninformed traders collectively can be referred to as D^U, which is a normally distributed random variable independent of V with mean zero and variance σ_u^2. If D^U is positive, then uninformed traders are buying on net. Conversely, uninformed traders are selling the asset on net, if D^U is negative. Though the informed trader knows V, he does not know D^U prior to submitting his orders. Effectively this precludes the informed trader from conditioning on the market-clearing price, as it is usual in a rational expectations model.

Once receiving these orders, the auctioneer determines P, the market clearing price. Kyle assumes free entry into the auctioneering market and therefore the auctioneer has no monopoly power, so that he earns zero profits and P is determined by the following equation:

$$P = E[V|D^I + D^U] .$$

To arrive at a value for P, the auctioneer only takes into account the sum of the orders issued by the informed trader and the uninformed traders: $D^I + D^U$. P depends on the sum of the orders because he cannot differentiate between the orders issued by the informed trader from the rest. Note that D^U is an exogenous variable, but D^I depends on the informed trader's trading strategy. The informed trader knows that his order has some effect on the price created by the auctioneer. Since he is risk neutral, the informed trader will seek to maximize his expected profit. He accomplishes this goal by considering each possible value of V and choosing the value of D^I that maximizes:

$$E[D^I(V - P)|V] .$$

These two equations illustrate that the auctioneer's strategy for setting the asset's price depends on D^I while the informed trader's strategy for setting D^I depends on his perceived effect of D^I on P.

Kyle first conjectures general functions for the pricing rule and the informed trader's demand, then he solves for the parameters assuming the informed trader maximizes his profits conditioning on his information set, i. e. $D^I = \operatorname{argmax} E[D^I(V - P)|V]$ and the market maker sets prices equal to $P = E[V|D^I + D^U]$.

Although the proof is not shown here, in equilibrium the market maker will choose a price such that

$$\mathrm{P} = \lambda(\mathrm{D}^I + \mathrm{D}^U)$$

and the informed trader will choose D^I such that

$$\mathrm{D}^I = \beta V$$

where λ and β are positive coefficients that solely depend on σ_v^2, the variance of V, the normally distributed random variable for the asset's payoff, and σ_u^2 the variance of

D^U, the normally distributed random variable for the uninformed traders' orders. The exact expressions (not derived here) for λ and β are:

$$\lambda = \frac{1}{2}\sqrt{\frac{\sigma_v^2}{\sigma_u^2}}$$

$$\beta = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}}\,.$$

If λ has a high value, then order flow has a high impact on prices, and we say that the particular asset is not very liquid. B, on the other hand, is rather low, which is interpreted as informed traders issuing less aggressive orders in an effort to minimize the impact of their own trades on price.

Empirical Market Microstructure

As transaction-by-transaction or high frequency data from a variety of sources has become available, empirical market microstructure has grown extensively. Most papers use high frequency data to predict transaction costs, estimate limit-order book models for intraday trading strategies, and estimate the liquidity of the market. There are a few papers, though, that do not estimate market microstructure models per se, but use high frequency data to answer questions relevant to the asset pricing field, corporate finance field, and economics in general. For example, Andersen et al. [6] use intraday data to obtain better measures of the volatility of asset prices, Chen, Goldstein and Jiang [13] estimate the PIN measure to answer questions relevant to corporate finance, and Andersen et al. [8] and [7] use intraday data to better identify the effect macroeconomic news announcements have on asset prices.

In what follows we describe the most commonly used empirical estimations of liquidity or adverse selection costs. The most general measure of adverse selection costs that does not assume a particular economic model is Hasbrouck [27]. He assumes the quote midpoint is the sum of two unobservable components,

$$q_t = m_t + s_t$$

where m_t is the efficient price, i. e., the expected security value conditional on all time-t public information, and s_t is a residual term that is assumed to incorporate transient microstructure effects such as inventory control effects, price discreteness, and other influences that cause the observed midquote to temporarily deviate from the efficient price. As such Hasbrouck [27] further assumes that $\mathrm{E}[\mathrm{s}_t] = 0$ and that it is covariance stationary which implies that microstructure imperfections do not cumulate over time, i. e., $\mathrm{E}_t[\mathrm{s}_{t+k}] \rightarrow \mathrm{E}[\mathrm{s}_{t+k}] = 0$ as $\rightarrow \infty$. The efficient price evolves as a random walk,

$$m_t = m_{t-1} + w_t \tag{1}$$

where $\mathrm{E}[\mathrm{w}_t] = 0$, $\mathrm{E}[\mathrm{w}_t^2] = \sigma_w^2$, $\mathrm{E}[\mathrm{w}_t \mathrm{w}_\tau] = 0$ for $t \neq \tau$ and w_t is also covariance stationary. The innovations, w_t, reflect updates to the public information set including the most recent trade. The market's signal of private information is the current trade innovation defined as $x_t - \mathrm{E}[x_t|\Phi_{t-1}]$, where Φ_{t-1} is the public information set at time $t-1$. The impact of the trade innovation on the efficient price innovation is $\mathrm{E}[\mathrm{w}_t|x_t - \mathrm{E}[x_t|\Phi_{t-1}]]$. Hence, two measures of information asymmetry, or trade informativeness, that Hasbrouck [27] proposes are:

$$\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]]) = \sigma_{w,x}^2$$

an absolute measure of trade informativeness and

$$R_w^2 = \frac{\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]])}{\mathrm{Var}(w_t)} = \frac{\sigma_{w,x}^2}{\sigma_w^2}$$

a relative measure of trade informativeness. The random walk decomposition, Eq. (1), on which these measures are based is unobservable. However, we can estimate $\sigma_{w,x}^2$ and σ_w^2 using a vector autoregressive (VAR) model,

$$r_t = \sum_{i=1}^{\infty} a_i r_{t-i} + \sum_{i=0}^{\infty} b_i x_{t-i} + v_{1,t}$$

$$x_t = \sum_{i=1}^{\infty} c_i r_{t-i} + \sum_{i=0}^{\infty} d_i x_{t-i} + v_{2,t}$$

where $r_t = q_t - q_{t-1}$ is the change in the quote midpoint, and x_t is an indicator variable that takes values $\{-1, +1\}$ whether the trade was seller-initiated or buyer-initiated according to the Lee and Ready [33] algorithm. Some papers also consider taking signed volume (number of transactions times shares traded) rather than signed transactions, but empirical evidence shows that what is important is the number of transactions not the number of shares traded.

Hasbrouck [27] estimates the VAR system using OLS. Wold's representation theorem states that any covariance-stationary process possesses a vector moving average (VMA) representation of infinite order, i. e. $\{r_t, x_t\}$ can be written as an infinite distributed lag of white noise, called the Wold representation or VMA. The minimum and maximum daily number of transactions among all the equities varies greatly, so the researcher has to set truncation points for each individual stock separately. Rather than use the Akaike and SIC information criteria to determine the optimal lag length, the purpose of the VAR estimation above is to get rid off all serial correlation. Once the lag lengths are set we can estimate the following VMA representation:

$$\begin{aligned} r_t &= \sum_{t=1}^{N} a_i^* v_{1,t-1} + \sum_{t=0}^{N} b_i^* v_{2,t-1} \\ x_t &= \sum_{t=1}^{N} c_i^* v_{1,t-1} + \sum_{t=0}^{N} \mathrm{d}_i^* v_{2,t-1} \, . \end{aligned}$$

Hence the trade-correlated component of the variance is equal to

$$\hat{\sigma}_{w,x}^2 = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) + \left(1 + \sum_{t=1}^{N} a_i^*\right)^2 \sigma_1^2 \, ,$$

where $\Omega = \mathrm{Var}(v_{1,t}, v_{2,t})$ and $\sigma_1^2 = \mathrm{Var}(v_{1,t})$ the variance of the random-walk component is

$$\hat{\sigma}_w^2 = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) \, .$$

Some Estimation Considerations

The VAR and VMA systems described above are not standard autoregressive models, in the sense that the index t is not a wall-clock index, but an event index, i. e., it is incremented whenever a trade occurs or a quote is revised. The choice between an event index and a wall-clock index depends on the goals of the analysis. If the analysis involves a single security, an event index is better than a wall-clock index because the process is more likely to be covariance stationary in event time than in wall-clock Hasbrouck (see p. 90 in [29]). However, when conducting a cross-sectional analysis or estimating a pooled regression, comparability across securities becomes the dominant consideration and one may want to adopt a wall-clock time in estimating the above equations.

Hasbrouck (see p. 39 in [29]) also points out that the overnight return will almost certainly have different properties than the intraday return and he suggests that one should drop the overnight return.

All told, researchers that use Hasbrouck's measure of adverse selection costs to test important economic hypothesis should feel very uncomfortable if their results depended on the way they estimate the VAR equations. As a robustness check researchers should estimate the VAR using different specifications, i. e., wall-clock time as opposed to event-time indexes, and researchers should sample quotes at different frequencies.

Madhavan, Richardson and Roomans Model

Similar to Hasbrouck [27], Madhavan, Richardson and Roomans [38] model the efficient price, m, as a random walk, in contrast they include an order flow innovation term,

$$m_t = m_{t-1} + \theta(x_t - E[x_t|x_{t-1}]) + \varepsilon_t \tag{2}$$

where θ measures the permanent price impact of order flow and ε_t is the public information innovation. The transaction price, p, is equal to the efficient price plus a stochastic rounding error term, ξ, and a market makers' cost per share of supplying liquidity, ϕ, i. e. compensation for order processing costs, inventory costs etc.

$$p_t = m_t + \phi x_t + \xi_t \tag{3}$$

Combining Eq. (2) and Eq. (3) we obtain,

$$m_t = p_t - p_{t-1} - (\phi + \theta)x_t + (\phi + \rho\theta)x_{t-1}$$

where ρ is the first-order autocorrelation of the signed trade variable, x_t. Then, the measure of permanent price impact, θ, alongside the temporary price impact of order flow, ϕ, the autocorrelation of signed trades, ρ, the unconditional probability that a transaction occurs within the quoted spread λ, and a constant, β, can be estimated using GMM applied to the following moment conditions:

$$E\begin{pmatrix} x_t x_{t-1} - x_{t-1}^2 \rho \\ |x_t| - (1-\lambda) \\ m_t - \beta \\ (m_t - \beta)x_t \\ (m_t - \beta)x_{t-1} \end{pmatrix} = 0$$

Future Directions

Hasbrouck [29] on page 7 lists a few outstanding significant questions in market microstructure. To this list we add two particularly important issues. First, the recent availability of good quality high frequency data has made it possible for researchers to answer a wide range of questions. This new data, though, also raises questions. In our opinion, it is imperative for researchers to determine under what circumstances more data is better. Some papers in the *realized volatility* literature have made headway in this direction by determining optimal sampling frequencies to estimate the volatility of assets with different liquidity. Future research should investigate what is the optimal frequency in estimating adverse selection costs and in event studies – studies that investigate the impact of public announcements on prices and trading in the hours surrounding its release. Second, most empirical and theoretical studies assume that trades affect prices, but prices do not affect trades (see, for example, Hasbrouck's VAR specification). Theory provides means of understanding why causality runs from trades to prices – trades are correlated with private information, so that trades cause asset price changes, with the underlying private information being the primitive cause. However, a more realistic setting is that in which there are heterogeneous beliefs and prices partially reveal other agent's information so that there is a learning process. Future research should relax the assumed exogeneity of trades. Finally, two productive areas of research are (i) the investigation of microstructure issues in fixed income markets, and (ii) studies that link microstructure to other areas in finance such as asset pricing and corporate finance.

Readings

Various economists have written books and articles about the field of market microstructure. This article is a short survey and here we compile an non-exhaustive list of publications that provide more comprehensive reviews of the literature: [12,25,29,35,37,41], and [42]. Martin Evans and Richard K Lyons have also written a useful manuscript entitled "Frequently Asked Questions About the Micro Approach to FX," even though its main focus is on foreign exchange markets it is also applicable to the market microstructure literature in general.

Bibliography

1. Akerlof GA (1970) The market for lemons: Quality uncertainty and the market mechanism. Q J Econ 84:488–500
2. Amihud Y, Mendelson H (1980) Dealership market: Market making with inventory. J Financ Econ 8:31–53
3. Amihud Y, Mendelson H (1987) Trading mechanisms and stock returns: An empirical investigation. J Financ 42:533–553
4. Amihud Y, Mendelson H (1991) Volatility, efficiency and trading: Evidence from the Japanese stock market. J Financ 46:1765–1790
5. Amihud Y, Mendelson H, Murgia M (1990) Stock market microstructure and return volatility: Evidence from Italy. J Bank Financ 14:423–440
6. Andersen T, Bollerslev T, Diebold FX, Labys P (2003) Modelling and forecasting realized volatility. Econometrica 71:579–626
7. Andersen T, Bollerslev T, Diebold FX, Vega C (2003) Micro effects of macro announcements: Real-time price discovery in foreign exchange. Am Econ Rev 93:38–62
8. Andersen T, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond, and foreign exchange markets. J Int Econ 73:251–277
9. Bagehot W (pseudonum for Jack Treynor) (1971) The only game in town. Financ Analysts J 27:12–14
10. Boehmer, Saar, Yu (2004) Lifting the veil: An analysis of pretrade transparency at the NYSE. J Financ 60:783–815
11. Brunnermeier MK (2001) Asset pricing under asymmetric information. Oxford University Press, Oxford
12. Bruno B, Glosten LR, Spatt C (2005) Market microstructure: A survey of microfundations, empirical results, and policy implications. J Financ Mark 8:217–264
13. Chen Q, Goldstein I, Jiang W (2007) Price Informativeness and Investment Sensitivity to Stock Price. Rev Financ Stud 20:619–650
14. Copeland T, Galai D (1983) Information effects and the bid-ask spread. J Financ 38:1457–1469
15. Diamond DW, Verrecchia RE (1991) Liquidity and the cost of capital. J Financ 46:1325–1359
16. Easley D, O'Hara M (1992) Time and the process of security price adjustment. J Financ 47:577–604
17. Easley D, Kiefer MN, O'Hara M, Paperman JB (1996) Liquidity, information, and infrequently traded stocks. J Financ 51:1405–1436
18. Easley D, Hvidkjaer S, O'Hara M (2002) Is information risk a determinant of asset returns? J Financ 57:2185–2221
19. Easley D, Engle RF, O'Hara M, Wu L (2008) Time varying arrival rates of informed and uninformed trades. J Financ Econ 6:171–207
20. Ellis K, Michaely R, O'Hara M (2000) The accuracy of trade classification rules: Evidence from NASDAQ. J Financ Quant Analysis 35:529–551
21. Fabozzi FJ, Fleming MJ (2004) In: Fabozzi FJ (ed) US treasury and agency securities, 6th edn. McGraw Hill, pp 175–196
22. Garman (1976) Market microstructure. J Financ Econ 3:257–275
23. Glosten L, Milgrom P (1985) Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. J Financ Econ 13:71–100
24. Goettler R, Parlour C, Rajan U (2007) Microstructure effects and asset pricing. Working Paper, UC Berkeley
25. Goodhart AE, O'Hara M (1997) High frequency data in financial markets: Issues and applications. J Empir Financ 4:74–114
26. Grossman S, Miller M (1988) Liquidity and market structure. J Financ 43:617–637
27. Hasbrouck J (1991) The summary informativeness of stock trades: An econometric analysis. Rev Financ Stud 4:571–595
28. Hasbrouck J (1992) Using the TORQ database. NYU Working Paper
29. Hasbrouck J (2007) Empirical market microstructure: The institutions, economics, and econometrics of securities trading. Oxford University Press, New York
30. Hendershott T, Seasholes M (2007) Market maker inventories and stock prices. Am Econ Rev Pap Proc 97:210–214
31. Ho T Stoll H (1983) The dynamics of dealer markets under competition. J Financ 38:1053–1074
32. Jensen MC (1978) Some anomalous evidence regarding market efficiency. J Financ Econ 6:95–102
33. Lee MC, Ready MJ (1991) Inferring trade direction from intraday data. J Financ 46:733–746
34. Lee CM, Radhakrishna B (2000) Inferring investor behavior: Evidence from TORQ data. J Financ Mark 3:83–111
35. Lyons R (2001) The microstructure approach to exchange rates. MIT Press, Cambridge

36. Madhavan (1992) Trading mechanisms in securities markets. J Financ 47:607–642
37. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–258
38. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. Rev Financ Stud 10:1035–1064
39. Milgrom P, Stokey N (1982) Information, trade, and common knowledge. J Econ Theor 26:17–27
40. Odders-White E (2000) On the occurrence and consequences of inaccurate trade classification. J Financ Mark 3:259–286
41. O'Hara M (1995) Market microstructure theory. Blackwell Publishers, Oxford
42. Parlour CA, Seppi DJ (2008) Limit order markets: A survey. In: Boot AWA, Thakor AV (eds) Handbook of financial intermediation and banking
43. Schwert WG (2003) Anomalies and market efficiency. In: Harris M, Stulz RM, Constantinides GM (eds) Handbook of the economics of finance, vol 15. North-Holland
44. Smidt (1971) The road to an efficient stock market. Financ Analysts J 27:18–20; pp 64–69
45. Stoll H, Whaley R (1990) Stock market structure and volatility. Rev Financ Stud 3:37–71
46. Vega C (2005) Stock price reaction to public and private information. J Financ Econ 82:103–133
47. Viswanathan S, Wang J (2004) Inter-dealer trading in financial markets. J Bus 77:49–75

Market Microstructure, Foreign Exchange

Carol Osler
Brandeis International Business School,
Brandeis University, Waltham, USA

Article Outline

Glossary

Barrier options Options that either come into existence or disappear when exchange rates cross pre-specified levels. Barriers can be triggered by price rises or declines and reaching a barrier can either extinguish or create an option. An "up-and-out call," for example, is a call option that disappears if the exchange rate rises above a certain level. A "down-and-in put," by contrast, is created if the exchange rate falls to a certain level.

Bid-ask spread The difference between the best (lowest) price at which one can buy an asset (the ask) and the best (highest) price at which one can sell it (the bid). In quote-driven markets both sides of the spread are set by one dealer. In order-driven markets, the "best bid and the best offer" (BBO) are likely to be set by different dealers at any point in time.

Brokers Intermediaries in the interbank foreign exchange market that match banks willing to buy with banks willing to sell at a given price. Two electronic brokerages – EBS (Electronic Broking Service) and Reuters – now dominate interbank trading in the major currencies. In other currencies voice brokers still play an important role.

Call markets Financial markets that clear periodically rather than continuously. During a specified time interval, agents submit orders listing how much they are willing to buy or sell at various prices. At the end of the interval a single price is chosen at which all trades will take place. The price is chosen to maximize the amount traded and is essentially the intersection of the supply and demand curves revealed by the submitted orders.

Clearing The administration process that ensures an individual trade actually takes place. The amounts and direction are confirmed by both parties and bank account information is exchanged.

Corporate (or commercial) customers One of the two main groups of end-users in the foreign exchange market. Includes large multinational corporations, middle-market corporations, and small corporations. Their demand is driven almost entirely by international trade in goods and services, since traders at these firms are typically not permitted to speculate spot and forward markets.

Covered interest arbitrage A form of riskless arbitrage involving the spot market, the forward market, and domestic and foreign deposits.

Dealership market See Quote-driven markets.

Delta-hedge A delta-hedge is designed to minimize first-order price risk in a given position. That is, small price changes should change the agent's overall position by only a minimal amount (ideally zero). A delta-hedge gets its name from an option's "delta," which is the first derivative of the option's price with respect to the price of the underlying asset. To delta-hedge a long call (put) option position, the agent takes a short (long) position in the underlying asset equal in size to the option's delta times the notional value of the option.

Expandable limit order An order whose quantity can be expanded if it is crossed with a market order for a larger quantity.

Financial customers One of the two main groups of end-users in the foreign exchange market. Includes hedge funds and other highly-leveraged investors, institutional investors such as mutual funds, pension funds, and endowments, multilateral financial institutions such as the World Bank or the IMF, broker-dealers, and regional banks.

Feedback trading The practice of trading in response to past returns. Positive-feedback trading refers to buying (selling) after positive (negative) returns. Negative-feedback trading refers to selling (buying) after positive (negative) returns.

Foreign exchange dealers Intermediaries in the foreign exchange market who stand ready, during trading hours, to provide liquidity to customers and other dealers by buying or selling currency. Salespeople manage relationships with clients; interbank traders manage the inventory generated by customer sales, and also speculate on an extremely high-frequency basis, by trading with other banks; proprietary traders speculate on a lower-frequency basis in currency and other markets.

Forward market Currencies traded in forward markets settle after more than two trading days (and infrequently after less than two trading days).

FX Foreign Exchange.

Limit order See "Order-driven markets."

Long position A long position arises when an agent owns an asset outright.

Market order See "Order-driven markets."

Order flow Buy-initiated transactions minus sell-initiated transactions over a given period. Since customers are always the initiators, their order flow is just customer purchases minus customer sales. In the interdealer market, a dealer initiates a trade if s/he places a market order with a broker or if the dealer calls out to another dealer.

Order-driven markets Also known as "limit-order markets." Asset markets in which participants can both supply liquidity or demand it, as they choose. Liquidity suppliers place limit orders, which specify an amount the agent is willing to trade, the direction, and the worst acceptable price. A limit buy order in the euro-dollar market, for example, might specify that the agent is willing to buy up to $2 million at $1.2345 or less. These limit orders are placed into a "limit-order book," where they remain until executed or canceled. Agents demanding liquidity place "market" orders, which state that the agent wishes to trade a specified amount immediately at whatever price is required to fulfill the trade. Market orders are executed against limit orders in the book, beginning with the best-priced limit order and, if necessary, moving to limit orders with successively less attractive prices. The foreign exchange interdealer markets for major currencies are dominated by two electronic limit-order markets, one run by EBS and the other run by Reuters.

Overconfidence A human tendency to have more confidence in oneself than is justified. Humans tend to overestimate their own personal and professional success ("hubris") and that they overestimate the accuracy of their judgments ("miscalibration").

Over-the-counter market See quote-driven market.

Picking-off risk The risk that a limit order will be executed against a better informed trader, leaving the limit-order trade with a loss.

Price-contingent orders Orders that instruct a dealer to transact a specified amount at market prices once a currency has traded at a pre-specified price. There are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate. Take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate.

Quote-driven markets Also known as "dealership markets" or "over-the-counter markets." An asset market in which dealers provide immediate liquidity to those needing it. During trading hours the dealers commit to trade at any time but at prices they quote. The price at which they are willing to buy, the "bid," is always no greater – and usually lower – than the price at which they are willing to sell, the "ask." Foreign exchange dealers transact with end-users in a quote-driven market.

Settlement The process by which funds actually change hands in the amounts and direction indicated by a trade.

Short position A short position arises when an agent sells an asset, possibly before actually owning the asset. A "short position in euros" could arise if a dealer starts with zero inventory and then sells euros. The dealer could keep the short euro inventory overnight, but will typically close the position out at the end of the trading day by buying the equivalent amount of euros. Note that the overall bank will not have a negative inventory position, since the bank maintains balances in every currency it trades. Someone "short euros in the forward market" would have entered into a forward contract to sell euros in the future.

Slippage The concurrent effect of a given trade on price.

Stop-loss orders See "Price-contingent orders."

Spot market Currencies traded in the spot market settle after two trading days (except for transactions between the US and Canadian dollars).

Swaps A swap in the foreign exchange market is analogous to a repo in the money market. One counterparty agrees to buy currency A in exchange for currency B from another counterparty in the spot market, and simultaneously agrees to sell currency A back to the same counterparty, and buy back currency B, at a future date. The spot transaction is at the spot rate, the forward transaction is at the forward rate.

Take-profit orders See "Price-contingent orders."

Technical Trading Trading based on technical analysis, an approach to forecasting asset-price movements that relies exclusively on historical prices and trading volume. In foreign exchange, the absence of frequent volume figures limits the information basis to past prices. Notably, technical forecasts do not rely on economic analysis. Nonetheless, many technical trading strategies have been demonstrated to be profitable in currency markets, even after considering transaction costs and risk.

Trading volume The value of transactions during a given time period.

Triangular arbitrage Between every three currencies A, B, and C there are three bilateral exchange rates. Triangular arbitrage is a way to make riskless profits if the A-per-B exchange rate does not equal the C-per-B exchange rate multiplied by the A-per-C exchange rate.

Definition of the Subject

"Foreign exchange microstructure" is the study of the currency trading process and high-frequency exchange-rate determination. The field is also called "the new microeconomics of exchange rates." Research in this area began in the late-1980s, when it became clear after many years of floating rates that traditional, macro-based exchange-rate models were not able to explain short-run dynamics. Research accelerated in the mid-1990s as currency trading systems became sufficiently automated to provide useful data.

Introduction

Foreign exchange microstructure research, or the study of the currency trading process, is primarily motivated by the need to understand exchange-rate dynamics at short horizons. Exchange rates are central to almost all international economic interactions – everything from international trade to speculation to exchange-rate policy. The dominant exchange-rate models of recent decades, meaning specifically the monetary model and the intertemporal optimizing models based on Obstfeld and Rogoff [149], come from macro tradition. These have some value relative to horizons of several years, but they have made little headway in explaining exchange rate dynamics at shorter horizons [69,116,134]. Shorter horizons are arguably of greater practical relevance.

As elucidated by Kuhn in his seminal analysis of scientific progress (1970), the emergence of major anomalies typically leads researchers to seek an alternative paradigm. Currency microstructure research embodies the search for a new paradigm for short-run exchange-rate dynamics.

The search for an alternative paradigm has focused on the currency trading process for a number of reasons. First, it is widely held that macroeconomic models are enhanced by rigorous "microfoundations" in which agent behavior is carefully and accurately represented. A rigorous microfoundation for exchange rates will require a thorough understanding of the currency trading process.

Researchers are also motivated to study currency trading by evident contradictions between the way currency markets actually work and the way exchange-rate determination is represented in macro-based models. As Charles Goodhart remarked of his time as adviser at the Bank of England, "I could not help but observe that some of the features of the foreign exchange … market did not seem to tally closely with current theory …" (p. 437 in [81]). To others with first-hand experience of the trading world, it seemed natural "to ask whether [the] empirical problems of the standard exchange-rate models … might be solved if the structure of foreign exchange markets was to be specified in a more realistic fashion" (p. 3 in [72]).

The emergence of currency-market research in recent years also reflects a confluence of forces within microstructure. By the mid-1990s, microstructure researchers had studied equity trading for over a decade, thereby creating a foundation of theory and a tradition of rigorous analysis. Meanwhile, technological advances at foreign-exchange dealing banks made it possible to access high-frequency transactions data. Currency markets – huge and hugely influential – were a logical new frontier for microstructure research.

Currency microstructure research – like all microstructure research – embodies the conviction that economic analysis should be based solidly on evidence. As articulated by Charles Goodhart, arguably the founder of this discipline, "economists cannot just rely on assumption and hypotheses about how speculators and other market agents may operate in theory, but should examine how they work in practice, by first-hand study of such markets" (p. 437 in [81]). Most papers in this area are empirical, and those that include theory almost always confront the theory with the data. The literature includes quite a few dealer surveys, reflecting a widespread appreciation of practitioner input. This survey, like the literature, emphasizes evidence.

Institutional Structure

This section describes the institutional structure of the foreign exchange market.

Basics

Foreign exchange trading is dispersed throughout the day and around the world. Active trading begins early in Asia, continues in Europe, peaks when both London and New York are open, and finally tapers off after London traders leave for the day. There is an "overnight" period during which trading is relatively thin, but it lasts only the few hours between the end of trading in London (around 19

GMT) and early trading in Sydney (around 22 GMT). In terms of geography, currency trading takes place in almost every big major city around the world, though there are major trading centers. These major centers are Singapore, Sydney, and Tokyo in Asia, London in Europe, and New York in North America.

Foreign exchange trading is an intensely competitive business. Price is one dimension of competition, but there are many others. When it evaluates trading institutions each year, *Euromoney* considers their pricing consistency, strategies and ideas for trading in options, and innovative hedging solutions [55]. Customer relations are also critically important. As in many industries, good customer relations are fostered by personal attention from salespeople and by perks for good customers, such as sports tickets and elegant feasts.

Unlike trading in stocks, bonds, and derivatives, trading in currency markets is essentially unregulated. There is no government-backed authority to define acceptable trading practices, nor is there a self-regulating body. Local banking authorities are limited to regulating the structure of trading operations: they typically require, for example, that clearing and settlement are administratively separate from trading. Any attempt to regulate trading itself, however, would encourage dealers to move elsewhere, an undesirable outcome since foreign exchange is an attractive industry – it pays high salaries and generates little pollution. In the absence of regulation, certain practices that are explicitly illegal in other markets, such as front-running, are not only legal but common in foreign exchange.

Market Size Spot and forward trading volume in all currencies is worth around $1.4 trillion per day [9]. If foreign exchange swap contracts are included, daily trading is roughly twice as large, at $3.2 trillion. By either figure, foreign exchange is the largest market in the world. Trading on the New York Stock Exchange (NYSE), for example, is on the order of $0.050 trillion per day [145], while daily trading in the US Treasury market, possibly the world's second-largest market, is on the order of $0.20 trillion [67]. Spot and forward trading, on which FX microstructure research has consistently focused, has grown rapidly for many years – average yearly growth since 1992 has been nine percent, and since 2004 has been 18 percent.

The vast bulk of foreign exchange trading involves fewer than ten currencies. The US dollar is traded most actively [9] due to its role as the market's "vehicle currency": to exchange almost any non-dollar currency for any other requires one to convert the first currency into dollars and then trade out of dollars into the second currency. The value of US dollars traded in spot and forward markets is roughly $1.2 trillion per day, over 86 percent of total traded value. Of course, two currencies are involved in every transaction so the total traded value every day is twice the day's trading volume. The euro accounts for 37 percent of all trading, a staggering $518 billion per day. The yen and the UK pound each account for a further sixteen percent of traded value. The next tier of currencies, comprising the Swiss franc, the Australian dollar and the Canadian dollar, accounts for eighteen percent of traded value. The remaining 150 or so of the world's convertible currencies account for merely thirty percent of traded value.

Only the dollar, the euro, and the yen are liquid throughout the trading day. Liquidity in most other currencies is concentrated during locally-relevant segments of the day. The Swedish krone, for example, is liquid only during European trading hours.

Quotation Conventions Each exchange rate is quoted according to market convention: dollar-yen is quoted as yen per dollar, euro-dollar is quoted as dollars per euro, etc. Trade sizes are always measured in units of the base (denominator) currency and the price is set in terms of the numerator currency. In euro-dollar, for example, where the euro is the base currency, a customer asking to trade "ten million" would be understood to mean ten million euros and the dealer's quotes would be understood to be dollars per euro. The minimum tick size is usually on the order of one basis point, though it is technically one "pip," meaning one unit of the fifth significant digit for the exchange rate as conventionally quoted. Examples will be more helpful: in euro-dollar, where the exchange rate is currently around $1.5000, one tick is $0.0001; for dollar-yen, where current exchange rates are roughly ¥ 110.00/$, one tick is ¥ 0.01.

The average trade size is on the order of $3 million [18]; trades of $50,000 or less are considered "tiny." Thus the average foreign exchange trade is roughly the same size as normal "block" (large) trades on the NYSE [125], which makes it large relative to the overall average NYSE trade. The average foreign exchange trade is smaller, however, than the average trade in the US Treasury market, where average interdealer trades vary from $6 to $22 million depending on maturity [67].

A Two-Tiered Market

The foreign exchange market has two segments or "tiers." In the first tier, dealers trade exclusively with customers. In the second tier, dealers trade primarily with each other. The interdealer market forms the market's core in the sense that customer prices are all based on the best available interdealer prices.

Interdealer trading in spot and forward markets now accounts for 38 percent of all trading [9]. This is down sharply from its 57 percent share in 1998, a change often ascribed to rapid consolidation in the industry. The current share is comparable to the share of interdealer trading on the London Stock Exchange, which was most recently estimated to be between 25 and 35 percent [163]. It is lower, however, than the share of interdealer trading in the US Treasury market, which was 68 percent in October 2007 [66].

The Customer Market The customer foreign exchange market is quote-driven, meaning that liquidity is provided by professional market makers. As in most such markets, currency dealers are under no formal obligation to provide liquidity, unlike specialists on the NYSE. Failing to provide liquidity on demand, however, could be costly to a dealer's reputation so dealers are extremely reliable. The market functioned smoothly even during the crisis of September 11, 2001. Spreads widened, as would be expected given the heightened uncertainty, but market makers stayed at their desks and trading continued uninterrupted [135].

The customer market is fairly opaque. Quotes and transactions are the private information of the two parties involved, the customer and the dealer. Unlike stock and bond markets, which publish trading volume daily, aggregate figures for customer trading volume are published only once every three years e. g.[9]. The lack of transparency is intensified by the tendency for large customer trades, meaning those over around $25 million, to be split into multiple smaller trades. Splitting trades, which is a way to minimize market impact and thus execution costs [16], also characterizes the London Stock Exchange [165], among other markets. Trade-splitting makes it more difficult for a dealer to know how much a customer actually intends to trade. Dealers like to know when customers are trading large amounts, since large trades move the market.

Dealers divide their customers into two main groups, and structure their sales force accordingly. The first group, financial customers, is dominated by asset managers but also includes non-dealing banks, central banks, and multilateral financial institutions. The asset managers, in turn, are divided into "leveraged investors," such as hedge funds and commodity trading associations (CTAs), and "real money funds," such as mutual funds, pension funds, and endowments. Financial customers account for 40 percent of foreign exchange trading [9], sharply higher than their 22 percent share in 1998 [9].

The second group of customers, referred to as "corporates," are commercial firms that purchase currency as part of ongoing real production activities or for financial purposes such as dividend payments or foreign direct investment. The share of such commercial trading has been steady at roughly twenty-percent for a decade [9]. Commercial customers tend to be the mainstay of profitability for smaller banks [136]. Financial customers, by contrast, tend to make bigger transactions and thus gravitate to bigger banks [154].

The customers listed above are all institutions. Unlike equity markets, where the trading of individuals for their own account can account for half of all trading, retail trading has historically been tiny in foreign exchange. The participation of individuals has been discouraged by large average trade sizes and by the need to establish lines of credit with dealing banks.

Though customer trading has historically been carried out over the telephone, trading over electronic communication networks is growing rapidly, spurred by the advent of new technologies [11]. Formal figures are not available, but dealers estimate informally that these new networks now account for over one fifth of all customer transactions. Major dealers run single-bank proprietary networks through which they are connected to individual customers. The biggest networks, however, are managed independently. Some of these multi-bank e-portals, such as FXAll, permit customers to get multiple quotes simultaneously. FXAll has appealed primarily to commercial customers, which have historically paid relatively wide spreads on average (as discussed later), since it has brought them enhanced pre-trade transparency, intensified competition among dealers and, according to dealers, smaller spreads. Other multi-bank e-portals, such as FXConnect or Hotspot FXi, focus on financial customers and are valued because they permit "straight-through processing" (STP), meaning fully automated clearing and settlement. STP handles back office functions far more efficiently than the traditional manual approach in part because it reduces the opportunity for human error. Another type of network, such as Oanda.com, target individuals trading for their own account, permitting them to trade with no more than a Paypal account. Though such retail trading has grown rapidly in the current century, dealers report that it does not yet affect market dynamics.

The Interdealer Market In the foreign exchange interbank market there are no designated liquidity providers. At every moment a dealing bank can choose whether to supply liquidity or demand it. A dealer needing liquidity can, of course, call another dealer and request a quote.

Until the mid-1990s such "direct dealing" accounted for roughly half of all interdealer trading [36], while the other half of interdealer trading was handled by voice brokers – essentially limit-order markets in which people match the orders. During this period the best indication of the market price was often indicative quotes posted on Reuters' "FXFX" screen.

The structure of interdealer trading changed dramatically after the introduction of electronic brokerages in 1992. In the major currencies, electronic brokerages not only took over from the voice brokers but also gained market share relative to direct dealing. Electronic Broking Service (EBS) now dominates in euro and yen while Reuters, the other major electronic brokerage, dominates in sterling. As the electronic brokerages took over, their best posted bid and offer quotes became the benchmark for market prices. By the end of the 1990s, voice brokers were important only in the "exotic" (relatively illiquid) currencies for which electronic brokers are unavailable. The speed of this transition reflects the intensity of competition in this market.

EBS and Reuters share a common, uncomplicated structure. Standard price-time priority applies. Hidden orders are not permitted. Limit orders are not expandable. Orders must be for integer amounts (in millions). Trading is anonymous in the sense that a counterparty's identity is revealed only when a trade is concluded. Dealers pay commissions on limit orders as well as market orders, though the commission on limit orders is smaller.

These markets have moderate pre- and post-trade transparency relative to most other limit-order markets. With respect to pre-trade information, price information is limited to the five best bid and offer quotes, and depth information is limited to total depth at the quotes unless it exceeds $20 million (which it usually does during active trading hours). The only post-trade information is a listing of transaction prices. The exchanges do not publish any trading volume figures.

Automated (program) trading on the electronic brokerages was introduced in 2004. Trading was restricted to dealers until 2006, but now certain hedge funds are permitted to trade on EBS. These shifts are reported to be a major source of the surge in trading between dealers and their financial customers since 2004 [9].

Objectives and Constraints

To construct exchange-rate models with well-specified microfoundations it is critical to know the objectives and constraints of major market participants. It is also critical to know the constraints that determine equilibrium.

Dealers' Objectives and Constraints Dealers are motivated by profits according to the conscious intent of their employers. Half or more of their annual compensation comes in the form of a bonus which depends heavily on their individual profits [153]. Profits are calculated daily and reviewed monthly by traders and their managers.

Dealers are constrained by position and loss limits which are, in turn, management's response to rogue trader risk, meaning the risk that traders will incur immense losses [43,81]. A single rogue trader can bring down an entire institution: Nick Leeson brought down Barings Bank in the early 1990s by losing $1.4 billion; John Rusnack brought down Allfirst Bank by losing $700 million. Such catastrophes could not occur in the absence of an information asymmetry that plagues every trading floor: management cannot know each trader's position at all times. Traders are technically required to record their profits and losses faithfully and in a timely manner, but as losses mount they sometimes resort to falsifying the trading record. Position- and loss-limits are intended to minimize the risk that losses mushroom to that point. Intraday position limits begin at around $5 million for junior traders, progress to around $50 million for proprietary traders, and can be far higher for executive managers. Data presented in Oberlechner and Osler [148] suggests that intraday limits average roughly $50 million. Overnight position limits are a fraction of intraday limits, and loss limits are a few percent of position limits.

Profit-maximization for dealers involves inventory management, speculation, and arbitrage. We review these activities in turn.

Inventory management Foreign exchange dealers manage their own individual inventory positions [18,81], tracking them in a "deal blotter" or on "position cards" [120]. Large dealers as well as small dealers typically choose to end the day "flat," meaning with zero inventory, and generally keep their inventory close to zero intraday as well. Average intraday inventory levels are $1 to $4 million in absolute value and account for less than five percent of daily trading activity [18,154]. Though these absolute levels far exceed the $0.1 million median inventory level of NYSE specialists [98], the NYSE inventories are much larger relative to daily trading (24 percent).

Dealers generally eliminate inventory positions quickly. The half-life of an inventory position is below five minutes for highly active dealers and below half an hour for less active dealers [18,155]. Fast inventory mean-reversion has also been documented for futures traders [130], but standard practice in other markets often differs markedly. On the NYSE, for example, the half-life of inventory averages over a week [127]. Even on the London Stock Exchange, which has an active interdealer market like foreign exchange, inventory half-lives average 2.5 trading days [85].

Foreign exchange dealers in the major currencies generally prefer to manage their inventory via interdealer trades, rather than waiting for customer calls. In consequence, recent studies of dealer practices find no evidence of inventory-based price shading to customers, e. g.[154]. This distinguishes currency dealers from those in some equity markets [127] and bond markets [51]. Currency dealers also do not shade prices to other dealers in response to inventory accumulation [18]. Instead, dealers wishing to eliminate inventory quickly choose more aggressive order strategies [18,154].

Speculation Foreign exchange dealers speculate actively in the interdealer market [81]. Indeed, according to a dealer cited in Cheung and Chinn [36], "[d]ealers make the majority of their profit on rate movement, not spread" (p. 447). Consistent with this, Bjønnes and Rime [18] find that speculative profits are the dominant source of dealer profitability at the good-sized bank they analyze. Dealers' speculative positions are based on information gathered from customers, from professional colleagues at other banks, and from real-time news services.

Arbitrage Some dealers also engage in arbitrage across markets, such as triangular arbitrage or covered interest arbitrage. The associated software originally just identified the arbitrage opportunities, but by now it can actually carry out the trades. Arbitrage opportunities, though typically short-lived, arise frequently and occasionally provide sizeable profits (2).

Customers' Objectives and Constraints The three main types of customers are active traders, meaning levered funds and proprietary traders; real-money funds; and commercial firms.

Active Currency Traders The objectives and constraints of active currency traders are in some ways consistent with those assigned to international investors in standard academic models. These groups are motivated by profits: proprietary traders are motivated by an annual bonus; hedge fund managers receive a share of the firm's net asset value growth in [169]. Further, their risk-taking is constrained since active currency traders, like dealers, face position limits. Notice, however, that active currency traders are not motivated by consumption and they do not care about consumption risk. Indeed, there is no reason to expect the objectives of financial market participants to be aligned with those of consumers. It is agency problems that drive a wedge between the objectives of consumers and traders in foreign exchange: the institutions that employ the traders have to align the traders' incentives with those of shareholders under conditions of asymmetric information, with the result that consumption is irrelevant. Agency problems have been shown to be of overwhelming importance in understanding financial management at corporations. It would appear risky to assume that agency problems do not exist at currency-management firms.

Active currency traders also differ, however, from the academic image of the international investor. The speculative horizons of active currency traders typically range from a day to a month – longer than a dealer's intraday horizon but still short by macro standards. Further, these traders rarely take positions in assets with fixed supplies, such as bonds or equities. Instead, they rely on forwards, other derivatives, or possibly deposits, which are in flexible supply. This seemingly simple observation may unlock a longstanding puzzle in international macro, the apparent irrelevance of bond supplies for exchange rates. Under the standard assumption that speculative agents invest in bonds (an asset with fixed supply) bond supplies should influence exchange rates. Since bonds are not widely used by active currency speculators, however, the irrelevance of bond supplies seems natural.

Common speculative strategies among active currency traders are based on (i) forward bias, (ii) anticipated trends or trend reversals, and (iii) anticipated macro news.

Real-Money Managers Most managers of real money funds do conform to the academic image of an international investor in terms of their investment horizon and their assets of choice: they take positions for a month or more and generally invest in bonds or equities. These managers do not, however, conform to that image in a separate, critical dimension: real-world real money managers generally ignore the currency component of their return. According to Taylor and Farstrup [178], who survey the currency management business,

> there are key participants in foreign exchange markets … that are not always seeking profit derived *from their currency positions*. … [I]n this category are international equity managers. While some managers factor in currency considerations as they go about picking foreign stocks, most are attempting to add value through stock, sector, and region bets rather than currency plays (p. 10, italics in original).

The decision not to forecast the currency component of returns is sometimes justified by pointing to the well-known inability of macro-based exchange-rate models to forecast more accurately than a random walk [134]. Further information about financial customers is presented in Sager and Taylor [169].

Note that all speculative positions are constrained in currency markets. In exchange-rate models this would be consistent with the assumption that speculators are risk averse. It would not, however, be consistent with the assumption that deviations from purchasing power parity or uncovered interest parity are instantaneously eliminated by infinite trading. This may help explain why macroeconomic evidence of long standing shows that these parity conditions do not hold over short-to-medium horizons.

Commercial Customers With only rare exceptions, commercial firms do not take overtly speculative positions in spot and forward foreign exchange markets. Goodhart [81] estimates that less than five percent of large corporate customers will speculate in the forward market, and dealers report that zero middle-market or small corporations speculate in that way. Indeed, many firms explicitly prohibit their trading staff – often administrators with other responsibilities besides trading – from engaging in such transactions. Rogue trader risk is one key motivation for this choice. To impede the deception that enables rogue trading, firms that permit speculation must "separate the front office from the back office," meaning they must prohibit traders from confirming or settling their own trades. This requires a separate staff to handle these functions [65]. The firms must also hire "compliance officers" to ensure that controls on the trading process are being observed faithfully (Federal Reserve Bank of New York, Best Practice 48). Since the vast majority of commercial firms need to trade only infrequently to carry out their real-side business, these heavy staffing requirements make speculative trading prohibitively expensive.

Another powerful reason why corporate customers avoid overt speculation is that it can raise corporate tax burdens. In the US, at least, profits from overtly speculative positions are accounted for differently from gains designed to offset losses on existing business exposures, with the result that speculative profits are taxed more heavily. If a treasurer wishes to speculate, s/he can do so at a lower cost by redistributing the firm's assets and liabilities around the world. Goodhart [81] lists additional reasons why corporate customers generally do not speculate in spot and forward markets.

The presence of non-financial customers provides a natural source of heterogeneity in the motivations for currency trading. Such heterogeneity is critical for modeling asset prices, and may thus be critical for the functioning of asset markets [142,143]. When all agents are rational speculators it is hard to find reasons why speculators would trade with each other. If the price is away from its fundamental value both agents should insist on taking the profitable side of any trade, which is impossible. If the price is at its equilibrium, however, there is no profit to be gained from trading.

In the foreign exchange market, commercial firms necessarily have different trading motivations from speculators. Speculative agents primarily care about currencies as a store of value and commercial traders primarily care about currencies as a medium of exchange. Thus the existence of high trading volumes is less difficult to explain in foreign exchange than in, say, equity markets. (In bond markets, an alternative trading motivation may be provided by insurers and others engaged in duration matching.)

To generate trading volume in models of equity markets, financial modelers typically introduce "liquidity traders" or "noise traders" [22,115], typically modeled as a pure random variable and verbally assigned some motivation for trading. For liquidity traders the motivation is exogenous portfolio rebalancing; for noise traders the motivation is often speculation based on misinformation [22]. Neither motivation is fully satisfactory to the profession, however. Portfolio rebalancing is not sufficient to account for observed trading volumes and the professional preference for assuming rationality is not well-served by the noise trader concept. In foreign exchange markets, commercial traders provide rational trading partners for rational speculators.

Constraints on Exchange Rates The institutional features outlined in this section reveal a key constraint on exchange rates. On most days the amount of currency purchased by end-users must (roughly) equal the amount sold by end-users. Though dealers stand ready to provide liquidity intraday, the fact that they generally go home flat means that the dealing community, as a whole, does not provide overnight liquidity. Within a day, the net purchases of any end-user group must ultimately be absorbed by the net sales of some other end-user group. The exchange rate is presumably the mechanism that adjusts to induce end-users to supply the required liquidity.

This same explicit constraint can be found in financial markets known as "call markets" (see glossary), where a single price is chosen to match the amount bought to the amount sold. Prominent call markets include the opening markets on the NYSE and the Paris Bourse.

The very real constraint that end-user purchases equal end-user sales over a trading day differs dramatically from the exchange-rate equilibrium condition common to standard macroeconomic models. That condition is, in essence, that money demand equals money supply. The evidence does not support the relevance of aggregate money demand/supply to day-to-day exchange-rate determination [153].

Intraday Dynamics

This section provides descriptive information about trading volume, volatility, and spreads on an intraday basis.

Intraday Patterns in Volume, Volatility, and Spreads

Trading volume, volatility, and interdealer spreads all vary according to strong intraday patterns that differ in certain key respects from corresponding patterns in bond and equity markets. Figure 1a and b shows these patterns for euro-dollar and dollar-yen, based on EBS trade and quote data over the period 1999–2001 [103].

As in other markets, trading volume (measured here by the number of interbank deals) and volatility move together. As Asian trading opens (around hour 22) they both rise modestly from overnight lows, after which they follow a crude U-shape pattern during Asian trading hours and then another U-shape during the London hours. They both peak for the day as London is closing and New York traders are having lunch and then decline almost monotonically, reaching their intraday low as Asian trading opens early in the New York evening.

Some back-of-the envelope figures may help make these trading-volume patterns concrete. In Ito and Hashimoto's 1999–2001 EBS database there were roughly eight trades per minute in euro-dollar and six trades in dollar-yen [103]. Together with the seasonal patterns, this suggests that overnight interdealer trading was on the order of one or fewer trades per minute while peak trading (outside of news events) was on the order of 10 (JPY) to 25 (EUR) trades per minute. Current interdealer trading activity would be substantially larger, reflecting subsequent market growth.

Bid-ask spreads almost perfectly mirror the pattern of volume and volatility. They are highest during the overnight period, and then decline as trading surges at the Asian open. As trading and volatility follow their double-U pattern during Asian and London trading hours, spreads follow the inverse pattern: they rise-then-fall during Asian trading and then rise-then-fall once again during the London morning. After London closes, spreads rise roughly monotonically to their overnight peaks.

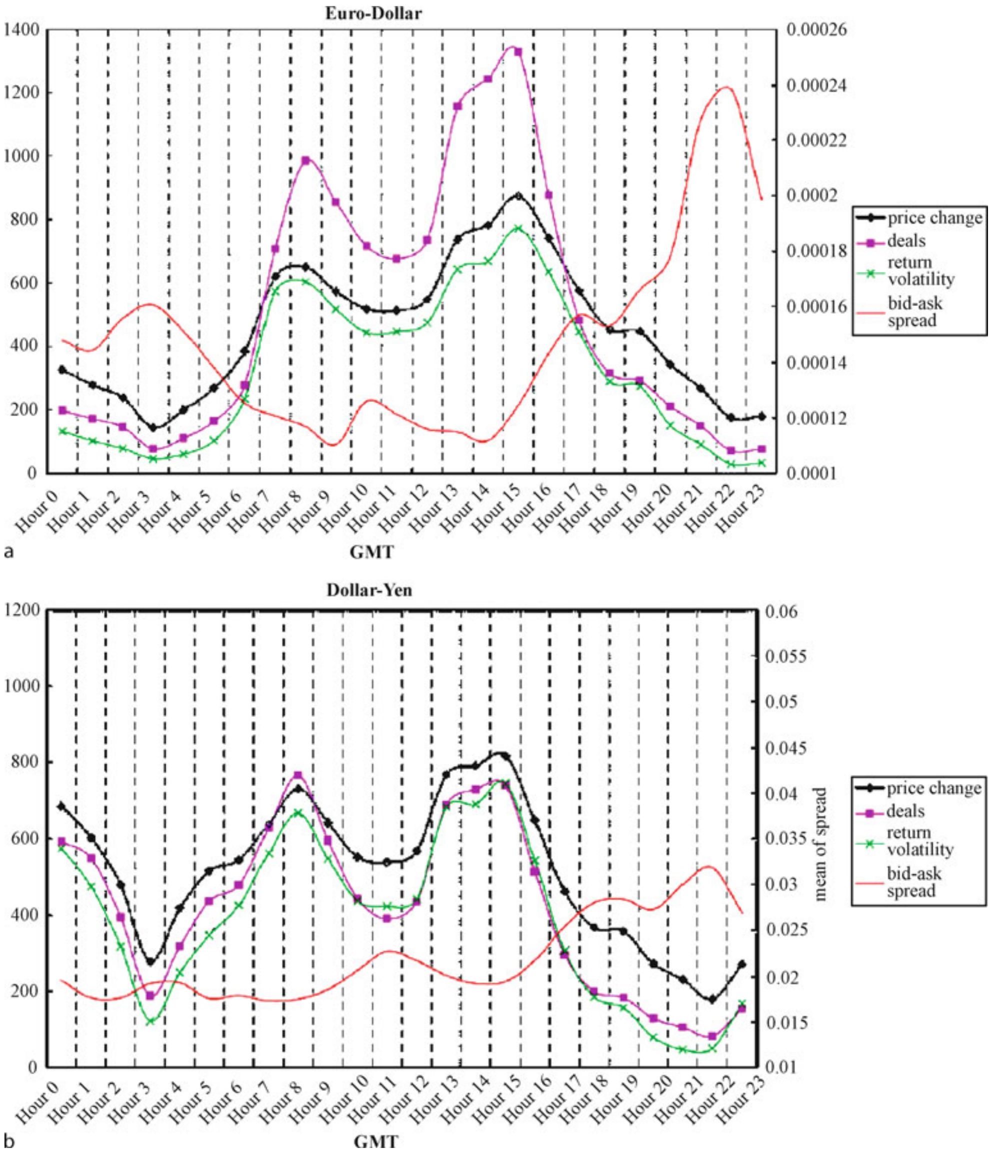
Conventional interdealer spreads, as reported in Cheung and Chinn [36], average three basis points in euro-dollar and dollar-yen, the two most active currency pairs. In sterling-dollar and dollar-swiss, the next two most active pairs, these averaged five basis points. Dealers in both the US [36] and the UK [37] report that the dominant determinant of spreads is the market norm. One important reason spreads widen is thin trading and a hectic market. Another important reason is market uncertainty [36], which is often associated with volatility. Since volatility also increases inventory risk, it makes sense that volatility and spreads have been shown to be positively related [23,92,105].

This tendency for interdealer spreads to move inversely from volume and volatility is consistent with predictions from two conceptual frameworks. Hartmann [92] explains the relationship in terms of fixed operating costs, such as the costs of maintaining a trading floor and of acquiring real-time information. When trading volume is high these costs can easily be covered with small spread, and vice versa, so long as the extra volume is dominated by uninformed traders. The same explanation could also apply at the intraday horizon.

Admati and Pfleiderer [1] develop an asymmetric information model consistent with some of the key properties just noted. In their model, discretionary uninformed traders (who can time their trades) choose to trade at one time since this brings low adverse selection costs to dealers and thus low spreads. The low spreads encourage informed traders to trade at the same time and the information they bring generates volatility. Overall, this model predicts that trading volume and volatility move in parallel and both move inversely with spreads, consistent with the patterns in major foreign exchange markets.

In most equity and bond markets, spreads move in parallel with trading volume and volatility, rather than inversely, with all three following an intraday (single) U-shape. Notably, a similar U-shape characterizes interdealer foreign exchange markets in smaller markets, such as Russia's electronic interdealer market for rubles, which only operate for a few hours every day [141]. In Taipei's interdealer market, which not only has fixed opening and closing times but also closes down for lunch, spreads follow a double-U-shape: they begin the day high, tumble quickly, and then rise somewhat just before lunch; after lunch they follow roughly the same pattern [76]. This contrast suggests that there is a connection between fixed trading hours and this U-shape for spreads.

Madhavan et al. [128] provide evidence that high spreads at the NYSE open reflect high adverse-selection risk, since information has accumulated overnight. High



Market Microstructure, Foreign Exchange, Figure 1
Intraday Patterns for Volume, Volatility, Spreads, and the Number of Price Changes. Figures are calculated from tick-by-tick EBS trade and quote data during winter months during 1999–2001. Seasonal patterns are only slightly different in summer. (Source: [103]). Greenwich Mean Time

spreads at the close, by contrast, reflect high inventory risk, according to their evidence, since dealers cannot trade until the market re-opens the next morning. In less-liquid foreign exchange markets, such as those for emerging market currencies, the overnight period is relatively long and there is little overnight liquidity, so similar patterns may arise. The failure of interdealer spreads in major currencies to follow the pattern observed in equity and bond markets need not imply, however, that adverse selection is irrelevant in the interdealer markets. In the major currencies, the overnight period is short and liquid (relative to other assets), so adverse-selection risk may not rise as sharply as the market opens and inventory risk may not rise as sharply as the overnight period approaches. In this case adverse selection could be relevant but subordinate to other factors, such as Hartmann's fixed operating costs.

Weekends are a different story, since foreign exchange trading largely ceases from about 21 GMT on Fridays until

21 GMT on Sundays. The previous analysis suggests that foreign exchange spreads might be particularly wide on Monday mornings in Tokyo and Friday afternoons in New York. There is support for the first of these implications: Ito and Hashimoto [103] provide tentative evidence that spreads are indeed exceptionally wide on Monday mornings in Tokyo.

Minute-by-minute data show that volume and volatility spike sharply at certain specific times of day [12]. In the New York morning there are spikes at 8:20, 8:30, 10 and 11 am, reflecting the opening of derivatives exchanges, the release of US macro news, standard option expiration times, and the WM/Reuters fixing (at 4 pm London time; this is a price at which many banks guarantee to trade with customers), respectively. Further spikes occur at 2 pm, and 8 pm New York time, reflecting the closing of derivatives exchanges and Japanese news releases, respectively. The timing of these spikes differs slightly in summer when daylight saving time is adopted in the UK and the US but not Japan.

The high trading that typically accompanies macro news releases represents a further dimension on which the markets differ from the features assumed in macro-based exchange-rate models. In macro-based models all agents have rational expectations and all information is public. The release of macro news causes everyone's expectations to be revised identically so the price moves instantly to reflect the new expectation without associated trading volume.

Feedback Trading

The data provide substantial evidence of both positive and negative feedback trading in foreign exchange. Sager and Taylor [169] find evidence for positive feedback trading in interdealer order flow using Granger-causality tests applied to the Evans and Lyons [58] daily data. Marsh and O'Rourke [131] and Bjønnes et al. [18] find evidence for negative feedback trading in semi-daily commercial-customer order flow but not in corresponding financial-customer order flow. Daniélsson and Love [44] find evidence of feedback trading in transaction-level interdealer trading data.

Feedback trading can greatly influence asset-price dynamics. For example, Delong et al. [45] show that in the presence of positive-feedback traders, the common presumption that rational speculators stabilize markets is turned on its head, and rational speculators intensify market booms and busts instead. Negative-feedback traders, by contrast, tend to dampen volatility.

There are at least three important sources of feedback trading in currency markets: technical trading, options hedging, and price-contingent orders. We discuss each in turn.

Technical Trading Technical trading is widespread in foreign exchange markets. Taylor and Allen [180] show that 90 percent of chief dealers in London rely on technical signals. Cheung and Chinn [36] find that technical trading best characterizes thirty percent of trading behavior among US dealers and the fraction has been rising. Similar evidence has emerged for Germany [137] and Hong Kong (Lui and Mole 1998).

Trend-following technical strategies generate positive-feedback trading. Froot and Ramadorai [74] present evidence for positive-feedback trading among institutional investors: their results indicate that, for major currencies vs. the dollar, a one standard deviation shock to current returns is associated with an 0.29-standard-deviation rise in institutional-investor order flow over the next thirty days.

Contrarian technical strategies generate negative feedback. For example, technical analysts claim that "support and resistance" levels are points at which trends are likely to stop or reverse, so one should sell (buy) after rates rise (fall) to a resistance (support) level. Support and resistance levels are a day-to-day topic of conversation among market participants, and most major dealing banks provide active customers with daily lists of support and resistance levels.

Option Hedging Option hedging also generates both positive- and negative-feedback trading. To illustrate, consider an agent who buys a call option on euros. If the intent is to speculate on volatility, the agent will minimize first-order price risk (delta-hedge) by opening a short euro position. Due to convexity in the relationship between option prices and exchange rates, the short hedge position must be modestly expanded (contracted) when the euro appreciates (depreciates). The dynamic adjustments therefore bring negative-feedback trading for the option holder and, by symmetry, positive-feedback trading for the option writer.

Barrier options – which either come into existence or disappear when exchange rates cross pre-specified levels – can trigger either positive- or negative-feedback trading and the trades can be huge. Consider an "up-and-out call," a call that disappears if the exchange rate rises above a certain level. If the option is delta-hedged it can trigger substantial positive-feedback trading when the barrier is crossed: since the short hedge position must be eliminated, the rising exchange rate brings purchases of the underlying asset. The entire hedge is eliminated all at once, however, so the hedge-elimination trade is far larger than the modest hedge adjustments associated with plain-vanilla options. Many market participants pay close attention to the levels at which barrier options have been written, and make efforts to find out what those levels are. Related option types, such as Target Resumption Notes (TARNs), also trigger substantial feedback trading but tend to spread it out.

Price-Contingent Orders Price-contingent customer orders are the third important source of feedback trading in foreign exchange. These are conditional market orders, in which the dealer is instructed to transact a specified amount at market prices once a trade takes place at a pre-specified exchange-rate level. There are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate, thereby generating positive-feedback trading. By contrast, take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate, thereby generating negative-feedback trading.

Take-profit orders are often used by non-financial customers that need to purchase or sell currency within a given period of time. Their option to wait is valuable due to the volatility of exchange rates. They can avoid costly monitoring of the market and still exploit their option by placing a take-profit order with a dealer. Financial customers also use take-profit orders in this way. Stop-loss orders, as their name implies, are sometimes used to ensure that losses on a given position do not exceed a certain limit. The limits are frequently set by traders' employers but can also be self-imposed to provide "discipline." Stop-loss orders can also be used to ensure that a position is opened in a timely manner if a trend develops quickly. Savaser [171] finds that stop-loss order placement intensifies prior to major macro news releases in the US.

One might imagine that these orders would tend to offset each other, since rising rates trigger stop-loss buys and take-profit sales, and vice versa. However, as discussed in Osler [151,152], differences between the clustering patterns of stop-loss and take-profit orders reduce the frequency of such offsets. Take-profit orders tend to cluster just on big round numbers: Stop-loss orders are less concentrated on the round numbers and more concentrated just beyond them (meaning above (below) the round number for stop-loss buy (sell) orders).

Since stop-loss and take-profit orders cluster at different points, offsets are limited and these orders create noticeable non-linearities in exchange-rate dynamics [151,152]. The presence of stop-loss orders, for example, substantially intensifies the exchange-rate's reaction to macro news releases [171]. Likewise, the tendency of take-profit orders to cluster at the round numbers increases the likelihood that trends reverse at such levels. This is consistent with the technical prediction, introduced earlier, that rates tend to reverse course at support; and resistance levels. Finally, the tendency of stop-loss orders to cluster just beyond the round numbers brings a tendency for exchange rates to trend rapidly once they cross round numbers. This is consistent with another technical prediction, that rates trend rapidly after a trading-range break out.

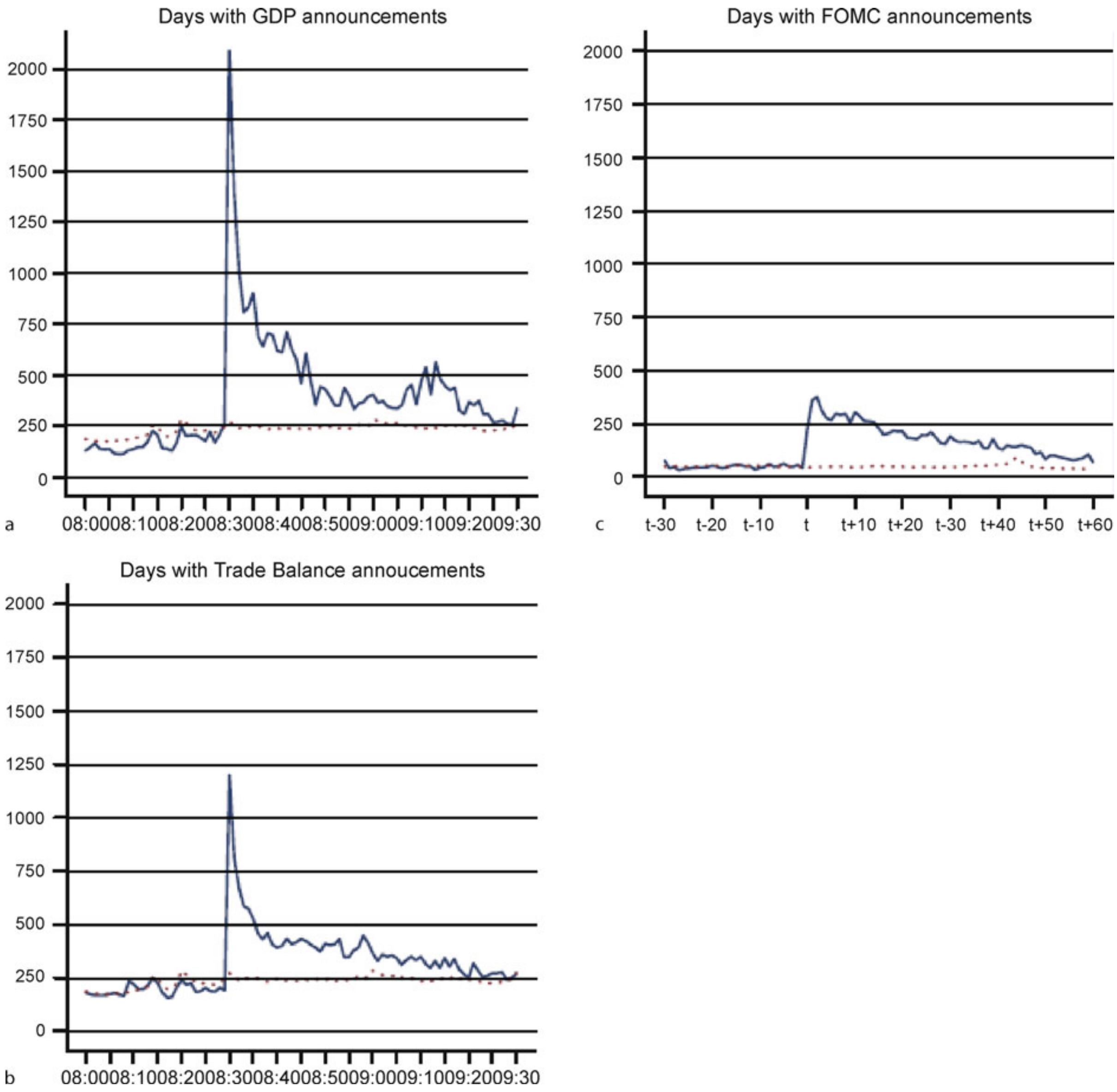
Market participants often report that stop-loss orders are responsible for fast intraday exchange-rate trends called "price cascades." In a downward cascade, for example, an initial price decline triggers stop-loss sell orders that in turn trigger further declines, which in turn trigger further stop-loss sell orders, etc. Upward cascades are equally possible: since every sale of one currency is the purchase of another, there are no short-sale constraints and market dynamics tend to be fairly symmetric in terms of direction (most notably, there is no equivalent to the leverage effect). Dealers report that price cascades happen relatively frequently – anywhere from once per week to many times per week. Osler [152] provides evidence consistent with the existence of such cascades.

News Announcements

Macro news announcements typically generate a quick surge in currency trading volume and volatility. As shown in Fig. 2a and b, which are taken from Chaboud et al. [31], volume initially surges within the first minute by an order of magnitude or more. Dealers assert that the bulk of the exchange-rate response to news is often complete within ten seconds [36].

Carlson and Lo [29] closely examines one macro announcement, the timing of which was unanticipated. They show that in the first half-minute spreads widened and in the second half-minute trading surged and the price moved rapidly. Chaboud et al. [31] shows that after the first minute volume drops back substantially, but not completely, in the next few minutes. The remaining extra volume then disappears slowly over the next hour. The response of returns to news is particularly intense after a period of high volatility or a series of big news surprises [48,54], conditions typically interpreted as heightened uncertainty.

The US macro statistical releases of greatest importance are the GDP, the unemployment rate, payroll employment, initial unemployment claims, durable goods orders, retail sales, the NAPM index, consumer confidence, and the trade balance [3]. Strikingly, money supply releases have little or no effect on exchange rates [3,25,



Market Microstructure, Foreign Exchange, Figure 2
Minute-by-minute trading volume, euro-dollar, around US scheduled macro news announcements. Based on tick-by-tick EBS trade data over 1999–2004. Trading volume relative to the intraday average. Source: [31]. Eastern Standard Time

36,62], consistent with the observation above that aggregate money supply and demand seem unimportant for short run exchange-rate dynamics.

Statistical releases bring a home-currency appreciation when they imply a strong home economy. A positive one-standard deviation surprise to US employment, which is released quite soon after the actual employment is realized, appreciates the dollar by 0.98 percent. For GDP, which is released with a greater lag, a positive one-standard deviation surprise tends to appreciate the dollar by 0.54 percent [3]. Responses are driven by associated anticipations of monetary policy: anything that implies a stronger economy or higher inflation leads investors to expect higher short-term interest rates [13] and thus triggers a dollar appreciation, and vice versa.

Federal Reserve announcements following FOMC meetings do not typically elicit sharp increases in trading volume and volatility [13]. Instead, FOMC announcements bring only a small rise in trading volume (Fig 2c) and tend to reduce exchange-rate volatility [33]. This suggests that Federal Reserve policy shifts are generally anticipated, which is encouraging since that institution prefers not to surprise markets.

Unanticipated changes in monetary policy do affect exchange rates. Fratscher [133] finds that an unanticipated 25 basis-point rise in US interest rates tends to appreciate the dollar by 4.2 percent. Kearns and Manners [108], who analyze other Anglophone countries, find that a surprise 25 basis-point interest-rate rise tends to appreciate the home currency by only 38 basis points. Kearns and Manners also note a more subtle dimension of response: If the policy shift is expected merely to accelerate an already-anticipated interest-rate hike, the exchange-rate effect is smaller (only 23 basis points, on average) than if the shift is expected to bring consistently higher interest rates over the next few months (43 basis points on average).

Evidence presented in Evans and Lyons [59] suggests that exchange rates overshoot in responses to news announcements. For some types of news, between a tenth and a quarter of the initial response is typically reversed over the four consecutive days. The reversals are most pronounced for US unemployment claims and the US trade balance. This contrasts strikingly with the well-documented tendency for the initial stock-price response to earnings announcements to be amplified after the first day, a phenomenon known as "post-earnings announcement drift" (Kothari [113] provides a survey). Nonetheless, over-reaction to fundamentals has been documented repeatedly for other financial assets [10,26,173].

Exchange-rate responses to a given macro news statistic can vary over time, as dealers are well aware [36]. During the early 1980s, for example, the dollar responded fairly strongly to money supply announcements which, as noted above, is no longer the case. This shift appears to have been rational since it reflected public changes in Federal Reserve behavior: in the early 1980s the Fed claimed to be targeting money supply growth, a policy it has since dropped. The possibility that such shifts are not entirely rational is explored in Bachetta and van Wincoop [7]. Cheung and Chinn [36] provide further discussion of how and why the market's focus shifts over time. Using daily data, Evans and Lyons [60] find little evidence of such shifting during the period 1993–1999. This could reflect the masking of such effects in their daily data or it could indicate that such shifting was modest during those years of consistent economic expansion and consistent monetary policy structure.

Information relevant to exchange rates comes from many more sources than macroeconomic statistical releases. Trading volume and volatility are triggered by official statements, changes in staffing for key government positions, news that demand for barrier options is rising or falling, reports of stop-loss trading, even rumors [48,147]. As documented in Dominguez and Panthaki [48], much of the news that affects the market is non-fundamental.

Numerous asymmetries have been documented in the responses to news. The effects of US macro announcements tend to be larger than the effect of non-US news [59,82]. Ehrmann and Fratzscher [54] attribute this asymmetry, at least in part, to the tendency for non-US macroeconomic statistical figures to be released at unscheduled times and with a greater lag. Ehrmann and Fratzscher also shows that exchange rates respond more to weak than strong European news, and Andersen et al. [3] report a similar pattern with respect to US announcements. This asymmetry is not well understood.

Carlson and Lo [29] shows that many interdealer limit orders are not withdrawn upon the advent of unexpected macro news. This might seem surprising, since by leaving the orders dealers seem to expose themselves to picking-off risk. It may not be the dealers themselves, however, that are thus exposed. The limit orders left in place may be intended to cover take-profit orders placed by customers, so the customer may be the one exposed to risk.

To be concrete: suppose a customer places a take-profit order to buy 5 at 140.50 when the market is at 140.60. The dealer can ensure that he fills the order at exactly the requested price by placing a limit order to buy 5 at 140.50 in the interdealer market. Suppose news is then released implying that the exchange rate should be 140.30. The dealer loses nothing by leaving the limit order in place: the customer still gets filled at the requested rate of 140.50.

This interpretation may appear to push the mystery back one step, because now the customer is buying currency at 140.50 when the market price of 140.30 would be more advantageous. Why wouldn't customers change their orders upon the news release, or withdraw them beforehand? This could reflect a rational response of customers to the high costs of monitoring the market intraday. Indeed, as noted earlier it is to avoid those costs that customers place orders in the first place. The Customers that choose not to monitor the market may not even be aware of the news.

Returns and Volatility

This section describes the basic statistical properties of returns and order flow.

Returns

Major exchange rates are often described as following a random walk, since it has long been well-documented

Market Microstructure, Foreign Exchange, Table 1
Autocorrelation of high-frequency returns. High-frequency autocorrelation of DEM returns, using Reuters indicative quotes over the period 1 October, 1992 through 30 September, 1993. Source: [33]

	5 min	10 min	15 min	30 min	Hourly
$\rho(1)$	–0.108	–0.093	–0.085	–0.066	–0.018
$\rho(2)$	–0.019	–0.030	–0.018	0.008	0.006
$\rho(3)$	–0.011	–0.002	0.006	0.024	–0.018

that daily returns to major exchange rates vis-à-vis the dollar are not autocorrelated and are almost entirely unpredictable. The random walk description is technically inaccurate, of course, since the variance of returns can indeed be forecast: it is statistically more accurate to describe the exchange rate as a martingale. (Further, at the highest frequencies returns are slightly negatively autocorrelated, as shown in Table 1 [33]). Whatever the nomenclature, the fact that current exchange rates provide better forecasts than standard fundamentals-based models [134] has long been a source of pessimism about exchange-rate theory in general.

Though the unconditional autocorrelation of daily returns is approximately zero, the conditional autocorrelation is not. Research has long shown that trend-following technical trading rules are profitable in major exchange rates [140]. Though returns to these rules seems to have declined in recent years, more subtle strategies remain profitable on a risk-adjusted basis [35]. Markov switching models also have predictive power for exchange rate returns [46,50], though the switching variables must include more than mean returns [117].

Daily returns are correlated across currencies, as one might expect given exchange-rate responses to news. The correlation between daily euro-dollar and sterling-dollar returns, for example, is 70 percent, while correlations between these European exchange rates and dollar-yen are smaller: both are 46 percent [12].

It has long been recognized that short-horizon exchange-rate returns are leptokurtotic. Kurtosis in euro-dollar returns, for example, is 24, 19, and 14 at the fifteen-minute, half-hour, and one hour horizons, respectively, all significantly higher than the level of three associated with the normal distribution [154]. Even at the two-day horizon kurtosis is still statistically significantly above three, though it has declined to five. These figures need not be constant. Osler and Savaser [154] demonstrate that a number of properties of price contingent orders impart high kurtosis to the distribution of returns. These properties include: high kurtosis in the orders' own size distribution, intraday seasonals in the execution of these orders; and the clustering patterns in their trigger rates described earlier. Stop-loss orders can also contribute to high kurtosis by contributing to price cascades. This analysis suggests that changes in market reliance on price-contingent orders could bring changes in the distribution of returns.

Within the overall distribution of returns there seems to have been a shift during the 1990s from the smallest returns, meaning those within one standard deviation of the mean, towards returns between one and five standard deviations [34]. The frequency of the most extreme returns, however, showed no trend.

Volatility

Unlike returns, volatility exhibits strong autocorrelation. As shown in Table 2, the first-order autocorrelation for daily volatility is typically above 0.50 and remains above 0.40 for at least a week. Evidence suggests that volatility is so persistent as to be fractionally integrated [12].

As recommended by Baillie and Bollerslev [8], volatility is typically captured with a GARCH(1,1) model or a close variant. Table 2b gives illustrative results from Chang and Taylor [33] showing that the AR component of the volatility process dominates (coefficients above 0.90) but the MA component is still significant. The MA component becomes increasingly important as the time horizon is shortened, though it remains subordinate. Table 2b also provides results suggesting that the double exponential distribution may fit return volatility better than the normal distribution. The thickness-of-tails parameter, "v," is two for the normal distribution but lower for the double exponential: estimates place it closer to unity than two.

Ederington and Lee [53] show, using 10-minute futures data for the DEM over July 3, 1989 through September 28, 1993, that the GARCH(1,1) model tends to underestimate the influence of the most recent shock and also shocks at long lags. These effects are captured better with an ARCH formulation that includes the lagged one-hour, one-day, and one-week return shock:

$$h_t = \alpha_0 + \sum_{i=1,6} \alpha_i \varepsilon^2_{t-i} + \alpha_7 \varepsilon^2_{\text{hour}} + \alpha_8 \varepsilon^2_{\text{day}} + \alpha_9 \varepsilon^2_{\text{week}} \,,$$

where h_t is estimated conditional volatility and ε_t is the shock to returns. These authors also find that daily and intraday seasonal patterns in volatility become fairly unimportant after controlling for announcements and ARCH effects. They conclude that "much of the time-of-day patterns and day-of-the-week patterns are due to announcement patterns" (p. 536).

Volatility usually rises upon news announcements, consistent with the analysis presented in III.C [53], but it can fall: Chang and Taylor [33] find that US Federal Re-

Market Microstructure, Foreign Exchange, Table 2

Strong autocorrelation in return volatility. a Daily realized volatilities constructed from five-minute returns based on Reuters indicative quote, July 1, 1987-December 31, 1993. Source: [160]. b Illustrative GARCH results assuming the normal distribution or the double-exponential distribution. Complete Reuters indicative quote for DEM, October 1992 through September 1993. Source: [33]

a	USD/DEM	USD/JPY	USD/GBP
$\rho(1)$	0.62	0.64	0.63
$\rho(2)$	0.52	0.53	0.54
$\rho(3)$	0.48	0.47	0.50
$\rho(4)$	0.45	0.44	0.47
$\rho(5)$	0.46	0.43	0.48

b	Hourly	30 Minutes	15 Minutes	5 Minutes
Normal Dist.				
α	0.045	0.035	0.098	0.100
	(3.83)	(4.36)	(8.32)	(13.82)
β	0.932	0.953	0.853	0.864
	(48.33)	(79.74)	(38.53)	(75.89)
Double-Exponential Dist.				
α	0.053	0.054	0.106	
	(5.07)	(4.86)	(4.97)	
β	0.930	0.936	0.878	
	(59.91)	(66.01)	(26.64)	
v	1.173	1.123	1.128	
	(41.71)	(52.14)	(58.82)	

serve news reduces volatility. This is consistent with the earlier finding that Fed news does not induce much extra trading. Volatility, like returns, can behave asymmetrically. Chang and Taylor [33] show that, during 1992, the volatility of dollar-mark was sensitive to US macro news but insensitive to German macro news. Such asymmetries need not be stable over time: Hashimoto [94] shows that asymmetries in the behavior of volatility changed dramatically around the Japanese bank failures of late 1997.

It is often hypothesized that volatility persistence derives from persistence in the flow of information, based on two premises: (i) volatility moves in parallel with trading volume, and (ii) trading volume is persistent because the advent of news is persistent. There is evidence to support both of these premises. Volatility and volume move together in most financial markets and foreign exchange is no exception, as shown in Fig. 1. Foreign exchange trading volume and volatility also move together at longer horizons [18,75]. Evidence also indicates persistence in the news process. Chang and Taylor [33], who count news releases on the Reuters real-time information system, find that autocorrelation in the number of news items is 0.29 at the one-hour horizon.

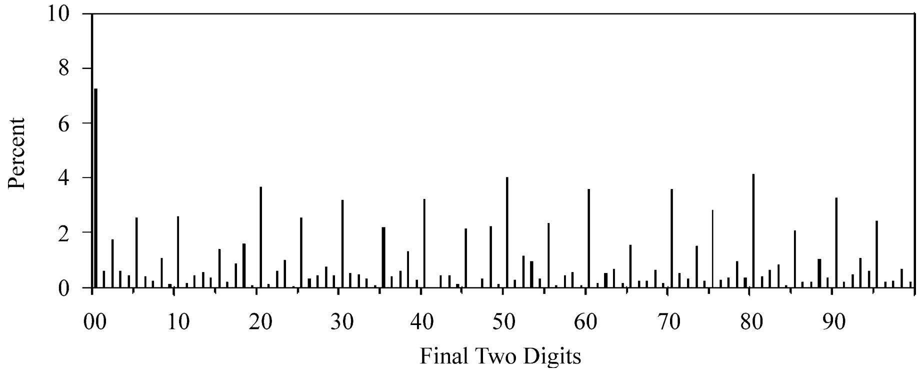
There is, however, little empirical evidence that directly traces volatility persistence in foreign exchange to news persistence. In fact, the only direct evidence on this point suggests that other factors are more important than news. Berger et al. [12] finds that persistence in news is primarily relevant to shorter-term volatility dynamics while long-run persistence in volatility is captured primarily by the low-frequency persistence in price impact, meaning the impact on exchange-rates of order flow. Figure 6, taken from Berger et al. [12], shows that daily price-impact coefficients for euro-dollar varied quite a bit during 1999–2004, and the series displays strong persistence at low frequencies. Further tests show that trading volume has modest explanatory power even after controlling for order flow.

Implied volatilities from exchange-traded options contracts have also been studied. Kim and Kim [111] find that implied volatilities in futures options are heavily influenced by volatility in the underlying futures price itself. They are not strongly influenced by news, and the few macro news releases that matter tend to reduce implied volatilities. Their analysis also indicates that implied volatilities tend to be lower on Mondays and higher on Wednesdays, though the pattern is not strong enough to generate arbitrage trading profits after transaction costs. Two studies show that daily volatility forecasts can be improved by using intraday returns information in addition to, or instead of, implied volatilities [132,160].

Order Flow and Exchange Rates, Part I: Liquidity and Inventories

Customer currency demand usually must net to around zero on trading days, as discussed earlier, and exchange-rate adjustment seems likely to be the mechanism that induces this outcome. If one group of customers decides to purchase foreign currency over the day, on net, the currency's value must rise to bring in the required liquidity supply from another group of customers. This implies, crudely, a relationship between net liquidity demand and exchange-rate returns.

To identify this relationship empirically one must distinguish liquidity-demand trades from liquidity-supply trades on a given day. We cannot simply look at trading volume or, equivalently, total buys or total sells, since it is the motivation behind the trades that matters. Instead we need to compare the purchases and sales of liquidity consumers. If they buy more than they sell then rates should rise to induce overnight liquidity supply and vice versa. The concept of "order flow" or, equivalently, "order im-



Market Microstructure, Foreign Exchange, Figure 3
Stop-loss and take-profit orders tend to be placed at round numbers. Data comprise the complete order book of the Royal Bank of Scotland in euro- dollar, sterling-dollar, and dollar-yen during the period September 1, 1999 through April 11, 2000. Chart shows the frequency with trigger rates ended in the 100 two-digit combinations from 00 to 99. Source: [151]

balances," which we examine next, can be viewed as a measure of net liquidity demand.

Interdealer Order Flow

In the interdealer market we identify liquidity demanders with either (i) those placing market orders or (ii) those calling other dealers to trade directly. When using transaction data from a broker, order flow is calculated as market buy orders minus market sell orders; when using direct dealing data, order flow is calculated as dealer-initiated buy trades minus dealer-initiated sell trades.

Evans and Lyons [58] were the first to show that interdealer order flow has substantial explanatory power for concurrent daily exchange-rate returns, a result that has been replicated in numerous studies [56,97]. Benchmark results are provided in Berger et al. [12], which has the advantage of a relatively long dataset. That paper shows that the raw correlation between daily returns and interdealer order flow is 65 percent for euro-dollar, 42 percent for sterling-dollar, and 49 percent for dollar-yen. Berger et al. estimates that an extra $1 billion in order flow in a given day appreciates the euro, the pound, and the yen by roughly 0.40 percent, with R^2s in the vicinity of 0.50. By contrast, it is well known that the explanatory power of standard fundamental variables is typically well below 0.10 [58].

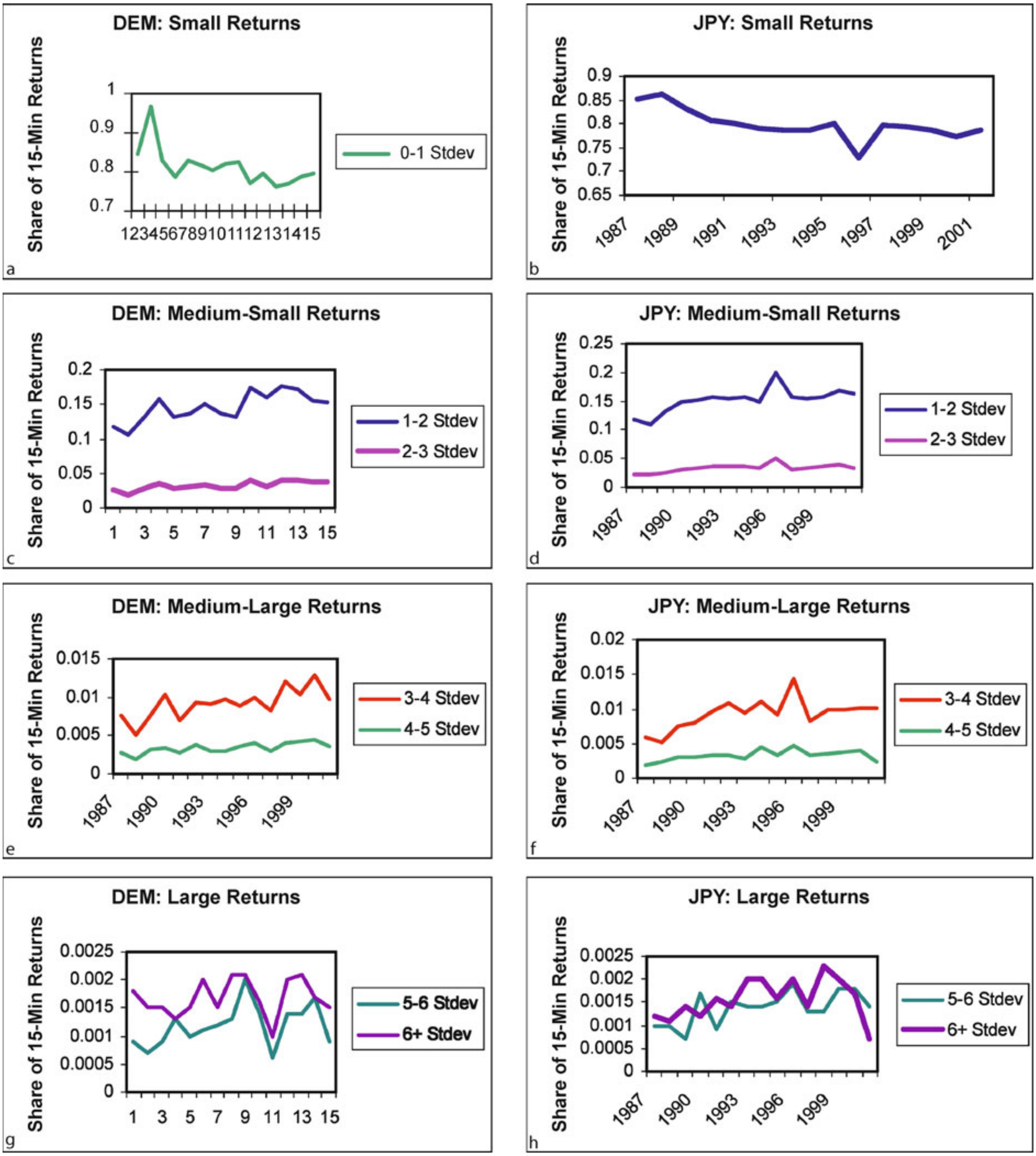
Evans and Lyons [58] and Rime, Sarno, and Sojli [166] find that the overall explanatory power of interdealer order flow for returns can be substantially increased by including order flow from other currencies. In Evans and Lyons [58], which uses daily interbank order flows for seven currencies against the dollar over four months in 1996, the joint explanatory power averages 65 percent and ranges as high as 78 percent.

Since feedback trading is ubiquitous in foreign exchange, one must consider the possibility that these correlations represent reverse causality – that returns are in fact driving order flow. Two studies investigate this possibility. Using daily data, Evans and Lyons [63] find that the influence of order flow on price survives intact after controlling for feedback effects; using transactions data, Daniélsson and Love [44] find that the estimated influence becomes even stronger after controlling for feedback trading.

Dealers have long recognized the importance of currency flows in driving exchange rates, and have said as much in surveys. In Gehrig and Menkhoff's survey [77], for example, over 86 percent of dealers said they rely on analysis of flows in carrying out their responsibilities. Indeed, the influence of order flow on exchange rates is a critical assumption in their trading strategies, as illustrated in the following debate over optimal management of stop-loss orders.

A dealer with a large stop-loss buy order could begin filling the order after the exchange-rate rises to the trigger price. Since the order-filling trades themselves will drive the price up, however, the average price paid will exceed the trigger rate, to the customer's disadvantage. The dealer could, alternatively, begin filling the order before the rate hits the trigger price. The buy trades will push the price up through the trigger rate and the average fill price will be closer to the trigger rate. The risk here is that the exchange rate bounces back down below the trigger rate, in which case the customer could justly complain of getting inappropriately "stopped out."

The key observation here is that the pros and cons of both strategy options are driven by the impact of order flow. Dealers do not view this as an hypothesis or as an assumption. To them it is something they know, in the same



Market Microstructure, Foreign Exchange, Figure 4
Frequency distribution of returns has shifted. Data comprise tick-by-tick Reuters indicative quotes over 1987–2001. Source: [34]

sense that one "knows" that the sun will disappear below the horizon at the end of the day (pace Hume). Dealers see order flow influence price too often and too consistently to question it.

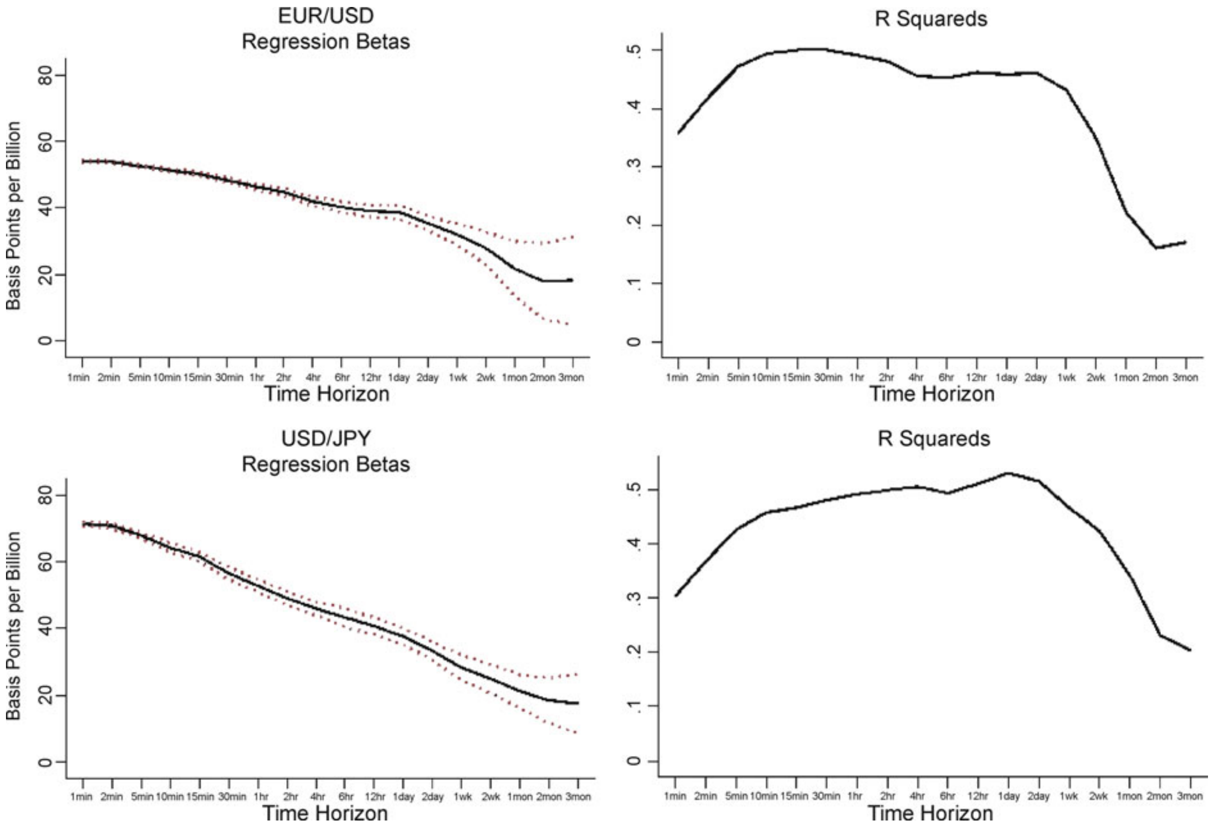
The estimated price impact of interdealer order flow varies according to order size, time of day, and time horizon. Price impact has a concave relationship to size [155], consistent with evidence from equity markets [93,104].

Market Microstructure, Foreign Exchange, Table 4
Autocorrelation coefficients for the number of exchange-rate relevant news items, 1 October 1992 through 30 September, 1993. Reuters News data. Source: [33]

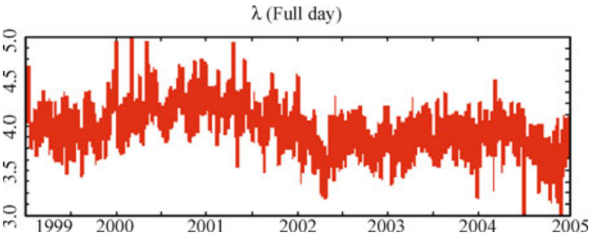
	Hourly	30 Min	15 Min	10 Min	5 Min
$\rho(1)$	0.27	0.22	0.34	0.09	0.06
$\rho(2)$	0.29	0.16	0.12	0.09	0.04
$\rho(3)$	0.22	0.15	0.11	0.08	0.05

This may reflect order splitting and other dealer strategies for minimizing the impact of large trades [17]. At the daily horizon, the price impact is linearly related to order flow, which makes sense since splitting a large trade into smaller individual transactions rarely takes more than a few hours. On an intraday basis, the price impact of interdealer order flow is inversely related to trading volume and volatility, as shown for dollar-yen in Fig. 7 [12]. As discussed earlier, spreads have a similarly inverse relation to trading volume and volatility (Fig. 1). This suggests, logically enough, that price impact is heavily influenced by spreads: when spreads widen, a given-sized transaction has a bigger price impact. Alternatively, however a third factor could be at work: depth. Depth presumably varies inversely with spreads and positively with trading volume intraday. Unfortunanely, information on depth is as yet almost nonexistent.

As time horizons lengthen the price impact of interdealer order flow declines monotonically [12]. For the euro, an extra $1 billion in order flow is estimated to bring an appreciation of 0.55 at the one-minute horizon but only 0.20 percent at the three-month horizon (Fig. 5, left). The explanatory power of interdealer order flow also varies with horizon but in a rising-falling pattern. The R^2 is 0.36 at the one-minute horizon, reaches 0.50 at the 30-minute horizon, stays fairly constant to the one-week horizon, and then falls sharply to about 0.17 percent at the two-month horizon (Fig. 5, right). Even at 17 percent, how-



Market Microstructure, Foreign Exchange, Figure 5
Response of returns to order flow at various horizons. Charts on the left show beta coefficients from regressions of returns on contemporaneous interdealer order flow for time horizons ranging from one minute to three months. Charts on the right show coefficients of determination from those same regressions. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. [12]



Market Microstructure, Foreign Exchange, Figure 6
Daily price impact coefficients for euro-dollar, 1999–2004. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. Source: [12]

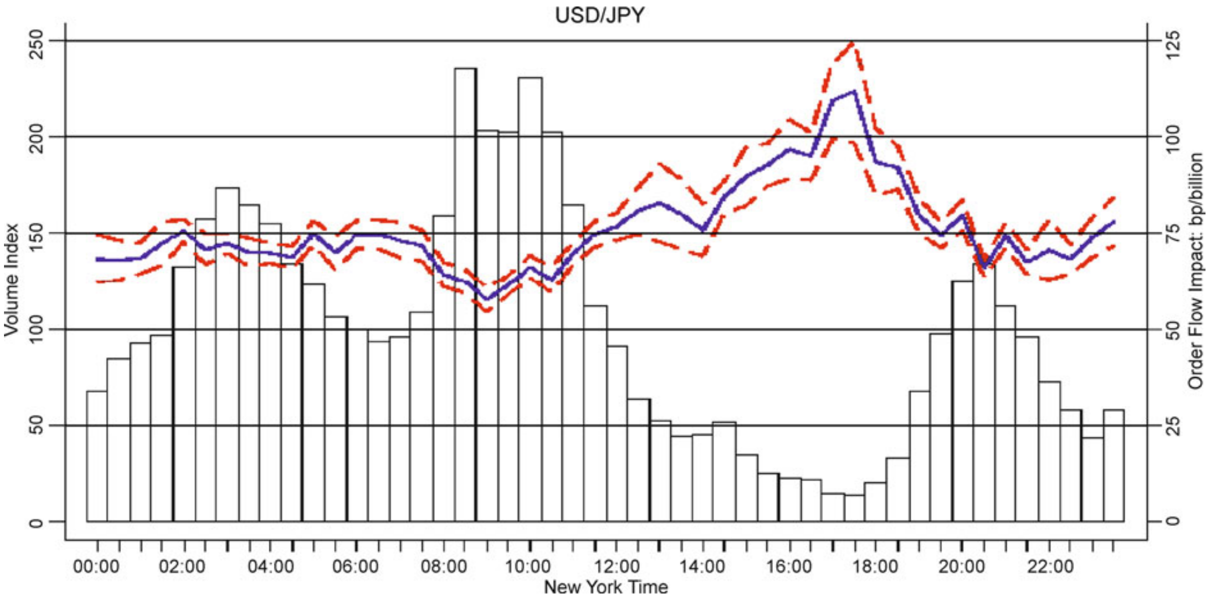
ever, the explanatory power of order flow at three months is substantially higher than has been achieved with other approaches. A similar pattern is found in Froot and Ramadorai, using institutional investor order flow, though they find a peak at roughly one month rather than one week [74]. They attribute the initial rise to positive-feedback trading.

The positive relation between interdealer order flow and exchange rates could be influenced by inventory effects as well as the liquidity effects described above. Inventory effects were, in fact, the first connection between order flow and asset prices to be analyzed in the broader microstructure literature, e. g. [177]. Dealers that provide liquidity to other dealers are left with an inventory position and thus inventory risk. Dealers charge a spread which compensates them for this risk. The spread, in itself, generates a positive relationship between order flow and returns: prices typically rise to the ask price upon buy orders and fall to the bid price upon sell orders.

Customer Order Flow

Order flow in the customer market is measured as customer-initiated buy trades minus customer-initiated sell trades. This is consistent with a liquidity interpretation on a trade-by-trade basis, since each customer effectively demands instantaneous liquidity from their dealer. Customer order flow, however, is not ideally suited to measuring customer net liquidity demand at daily or longer horizons. If a customer is coming to the market in response to an exchange-rate change, then the customer may be demanding liquidity from its own dealer at that instant while effectively supplying liquidity to the overall market.

This distinction proves critical when interpreting the empirical relation between daily customer order flow and exchange rates. There should be a positive relation between daily order flow and returns for customer groups



Market Microstructure, Foreign Exchange, Figure 7
Intraday Regression Betas and Average Trading Volume. Figure is based on the following regression: $\Delta s_t = \alpha + \beta \text{OF}_t + \eta_t$, where Δs_t is the return and OF_t is contemporaneous order flow. Regressions based on one-minute EBS trade data from 1999–2004 are run separately for each half hour of the trading day. Line shows estimated coefficients with standard error bands. Bars show order flow measured relative to the days' average (day's average set at 100). Source: [12]

that typically demand overnight liquidity. An increase in their demand for foreign currency, for example, should induce a rise in the value of foreign currency to elicit the required overnight supply. Implicit in that story, however, is a *negative* relation between order flow and returns for customer groups that typically supply overnight liquidity.

Researchers have documented repeatedly that, at the daily horizon, financial-customer order flow is positively related to returns while commercial-customer order flow is negatively related to returns. Confirming evidence is found in Lyons' [122] study of monthly customer order flows at Citibank; in Evans and Lyons [61] study of daily and weekly customer flows at the same bank; in Marsh and O'Rourke's [131] analysis of daily customer data from the Royal Bank of Scotland, another large dealing bank; and in Bjønnes et al. [18] comprehensive study of trading in Swedish kroner, and in Osler et al.'s [154] study of a single dealer at a medium-sized bank. The pattern is typically examined using cointegration analysis where the key relationship is between exchange-rate levels and cumulative order flow.

This pattern suggests that financial customers are typically net consumers of overnight liquidity while commercial customers are typically net suppliers. More direct evidence that commercial customers effectively supply overnight liquidity, on average, comes from evidence that commercial-customer order flow responds to lagged returns, rising in response to lower prices and vice versa. Marsh and O'Rourke [131] show this with daily data from the Royal Bank of Scotland. Bjønnes et al. [18] show this using comprehensive trading data on the Swedish krone sampled twice daily.

It is easy to understand why financial customers would demand liquidity: presumably they are speculating on future returns based on some information that is independent of past returns. Indeed, the identification of financial customers with speculation is explicit in Klitgaard and Weir's [112] study of currency futures markets. The IMM requires the agents they deem large speculators to report their positions on a weekly basis. Klitgaard and Weir show that their weekly position-changes are strongly correlated with concurrent exchange-rate returns. "[B]y knowing the actions of futures market speculators over a given week, an observer would have a 75 percent likelihood of correctly guessing an exchange-rate's direction over that same week" (p. 17).

It is not so immediately obvious why commercial customers would supply overnight liquidity, since our first image of a liquidity supplier is a dealer. Dealers supply intraday liquidity knowingly and are effectively passive in their trades with customers. By contrast, commercial customers are not supplying liquidity either knowingly or passively.

Commercial customers are, instead, just responding to changes in relative prices in order to maximize profits from their core real-side businesses. Suppose the foreign currency depreciates. Domestic firms note that their foreign inputs are less expensive relative to domestic inputs and respond by importing more, raising their demand for the foreign currency. This effect, a staple of all international economic analysis, has been well-documented empirically at horizons of a quarter or longer, e. g. [5]. On an intraday basis this effect is often evident in the behavior of Japanese exporting firms, which hire professional traders to manage their vast dollar revenues. These traders monitor the market intraday, selling dollars whenever the price is attractive. The vast majority of commercial customers need to buy or sell currency only occasionally so they can't justify hiring professional traders. They can use take-profit orders, however, to achieve the same goal, since this effectively enlists their dealers to monitor the market for them. At the Royal Bank of Scotland take-profit orders are 75 (83) percent of price-contingent orders placed by large corporations (middle-market) corporations [155], but only 53 percent of price-contingent orders overall.

The evidence to date suggests the following crude portrait of day-to-day liquidity provision in foreign exchange (a portrait first articulated in [18]). Financial customers tend to demand liquidity from their dealers, who supply it on an intraday basis. The dealing community as a whole, however, does not provide overnight liquidity. Instead, commercial customers supply the required overnight liquidity, drawn to the market by new, more attractive prices. Sager and Taylor [169] distinguish between "push" customers, who demand liquidity, and "pull" customers, who respond to price changes by providing liquidity. The market structure just outlined effectively identifies financial customers as short-run push customers and commercial customers as short-run pull customers.

This picture is extremely preliminary and will doubtless change as new evidence arrives. There is, for example, no theoretical or institutional reason why commercial customers must exclusively supply overnight liquidity or financial customers exclusively demand it. To the contrary, there are good theoretical reasons why the roles could sometimes be reversed. A change in commercial currency demand could result from forces outside the currency market, such as a war-induced rise in domestic economic activity, rather than a response to previous exchange-rate changes. In this case commercial end-users would consume liquidity rather than supplying it.

Market Microstructure, Foreign Exchange, Table 5

Order flow carries information about exchange-rate fundamentals. The table shows the R^2 statistics and associated marginal significance levels for the ability of daily customer order flow at Citibank during the period 1994 to 2001 to forecast upcoming announcements of key macro variables. Source: [57]

	US Output Growth				German Output Growth			
Forecasting Variables	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.
Output	0.002	0.003	0.022	0.092	0.004	0.063	0.069	0.006
	(0.607)	(0.555)	(0.130)	(0.087)	(0.295)	(0.006)	(0.009)	(0.614)
Spot Rate	0.001	0.005	0.005	0.007	0.058	0.029	0.003	0.024
	(0.730)	(0.508)	(0.644)	(0.650)	(0.002)	(0.081)	(0.625)	(0.536)
Order Flows	0.032	0.080	0.189	0.246	0.012	0.085	0.075	0.306
	(0.357)	(0.145)	(0.002)	(0.000)	(0.806)	(0.227)	(0.299)	(0.000)
All	0.052	0.086	0.199	0.420	0.087	0.165	0.156	0.324
	(0.383)	(0.195)	(0.011)	(0.000)	(0.021)	(0.037)	(0.130)	(0.000)

Speculative demand could also respond to changes in exchange-rate levels. Indeed, rational speculators are the *only* overnight liquidity suppliers in the widely-respected Evans and Lyons [58] model. In these models the trading day begins when agents arrive with arbitrary liquidity demands. The agents trade with their dealers, leaving the dealers with unwanted inventory. Dealers then trade with each other, redistributing their aggregate inventory but not reducing it. At the end of the trading day dealers sell the unwanted inventory to rational investors who are induced to supply the required liquidity by a change in the exchange rate. If the initial liquidity demanders have sold foreign currency, for example, the currency's value declines thus raising the risk premium associated with holding the currency. This encourages the risk-averse investors to take bigger positions in foreign assets, and as they enact the portfolio shift financial order flow is positive.

The Evans–Lyons scenario is necessarily simple. In a model with many assets, negative-feedback trading among financial customers requires that the currency has no perfect substitutes [88]. This condition holds in foreign exchange since exchange rates generally have low correlation with each other and with equities. For the negative feedback trading to be finite it is also required that speculators are risk-averse and/or face constraints on their trading. Though currency speculators appear to have a fairly high risk tolerance, their trading is always administratively constrained, as discussed earlier. The prevalence of contrarian technical trading strategies, such as those based on support and resistance levels, provides a further reason to expect negative-feedback trading among financial customers.

Despite these reasons to expect negative-feedback trading among financial customers, the evidence for it is thin and mixed. Financial agents do place a hefty share of take-profit orders [155], so a liquidity response from them is a fact. But their liquidity response may not be substantial relative to the overall market. Bjønnes et al. [18] study of trade in Swedish kroner and Marsh and O'Rourke's [131] study of customer trades at the Royal Bank of Scotland both find no sensitivity of financial order flow to lagged returns.

The influence of order flow on exchange rates described in this section works through liquidity effects. The broader microstructure literature refers to this influence in terms of "downward-sloping demand," highlighting that the demand for the asset has finite, rather than infinite, elasticity. Downward-sloping demand could explain why Froot and Ramadorai [74] find that the initial influence of institutional investor order flow disappears after roughly a year. Institutional investors – indeed, all speculative agents – have to liquidate positions to realize profits. When the positions are initially opened, the associated order flow could move the exchange rate in one direction; when the positions are liquidated the reverse order flow could move the exchange rate in the reverse direction.

Finite elasticity of demand is the underlying reason for exchange-rate movements in Hau and Rey's [96] model of equity and currency markets. Carlson et al. [30] develop a related exchange-rate model in which financial and commercial traders can be both liquidity suppliers and liquidity demanders. This model, which takes its critical structural assumptions directly from the microstructure evidence, predicts that financial (commercial) order flow is positively (negatively) related to concurrent returns, consistent with the evidence. It also predicts that these relations are reversed in the long run, consistent with evidence in Fan and Lyons [64] and Froot and Ramadorai [74]. Investors in the model have no long-run effect on exchange rates because they ultimately liquidate all their positions. Since commercial agents dominate long-run exchange rates, fundamentals such as prices and economic activity are important in the long run even though the may not dominate in the short run. In addition to being consistent with the microstructure evidence, this model is also consistent with most of the major puzzles in international macroeconomics, including: the apparent disconnect between exchange-rates and fundamentals, the increase in real-exchange-rate volatility upon the advent of floating rates, the short-run failure and long-run relevance of purchasing power parity, and the short-run failure of uncovered interest parity.

Order Flow and Exchange Rates

The influence of order flow on exchange rates is another aspect of the foreign exchange market that "does not seem to tally closely with current theory … " [81]. The equilibrium exchange rate in standard models adjusts to ensure that domestic and foreign money supplies equal corresponding money demands. The currency purchases or sales that accompany portfolio adjustments are not modeled and are considered unimportant. Indeed, order flow per se cannot be calculated in these models since they assume continuous purchasing power parity and/or continuous uncovered interest parity.

The contrast between microstructural reality and standard models is especially clear when we examine the mechanism through which news affects exchange rates. In macro-based models, the public release of information generates an immediate revision of shared expectations of future exchange rates, which in turn brings an immediate exchange-rate adjustment that requires no trading. Trading is unlikely, in fact, since no rational speculator would trade at any other price. Thus order flow in these models has no role in the exchange-rate adjustment to news.

The evidence shows, however, that order flow is the main conduit through which news influences exchange rates. Roughly two thirds of the influence of news on exchange-rate levels and volatility comes from the associated order flow [63,118]. During the "once-in-a-generation yen volatility" of 1998, "order flow [was the] most important … source of volatility," according to the investigation of Cai et al. [25], even more important than news and central bank intervention.

Reassuringly, the idea that order flow affects exchange rates is a natural extension of an important lesson learned after the advent of floating rates in the 1970s.

> [E]xchange rates should be viewed as prices of durable assets determined in organized markets (like stock and commodity exchanges) in which current prices reflect the market's expectations concerning present and future economic conditions relevant for determining the appropriate values of these durable assets, and in which price changes are largely unpredictable and reflect primarily new information that alters expectations concerning these present and future economic conditions (p. 726 in [73]).

There has long been extensive evidence that order flow influences price in stock markets [38,101,174]. In bond markets the evidence emerged later, due to constraints on data availability, but is nonetheless substantial [24,67,106,156,175,176]. Since exchange rates are asset prices they should be determined like other asset prices and thus order flow should be influential.

Order Flow and Exchange Rates, Part II: Information

So far we have considered two reasons why order flow could affect exchange rates: liquidity effects and inventory risk. This section considers a third and critically important reason: order flow carries private information.

The information hypothesis is suggested by evidence showing that much of the exchange-rate response to order flow is permanent. Payne [157], who decomposes returns into permanent and transitory components consistent with Hasbrouck [93], finds that "the permanent component accounts for … one quarter of all return variation" (p. 324). A permanent effect is implicit in Evans and Lyons' [58] evidence that order flow has strong explanatory power for daily exchange-rate returns, since daily returns are well described as a random walk. A permanent relation is also suggested by the finding, noted earlier, that cumulative order flow is cointegrated with exchange rates [18,110]. A permanent relation between order flow and price is not consistent with the inventory analysis presented earlier. A permanent relation is consistent with liquidity effects if the shifts in liquidity demand or supply are permanent. A permanent relation is inevitable, however, if order flow carries private fundamental information.

The influence of private fundamental information on asset prices was originally analyzed in equity-inspired models [79,115], which begin with the observation that sometimes customers often have private information about an asset's true value that dealers do not share. Since an informed customer only buys (sells) when the dealer's price is too low (high), dealers typically lose when they trade with such customers. To protect themselves from this adverse selection, dealers charge a bid-ask spread, ensuring that profits gained from trading with uninformed customers balance the inevitable losses from trading with informed customers [39]. Rational dealers ensure that their prices reflect the information communicated by a customer's choice to buy or sell [52,79]. Prices are "regret-free" in the sense that a dealer would not wish s/he had charged a higher (lower) price after learning that the customer wishes to buy (sell). Due to the spread, prices rise when informed customers buy and fall when informed customers sell. Meanwhile, others update their conditional expectation of the asset's true value and adjust their trades and quotes accordingly. Ultimately the information becomes fully impounded in price. Since the information is fundamental, the effect is permanent.

Types of Information

Private fundamental information in the foreign exchange market is likely to be structurally different from private fundamental information in a stock market. The fundamental determinants of a firm's value include many factors about which there can naturally be private information, such as management quality, product quality, and a competitor's strength. The fundamental determinants of a currency's value, by contrast are macroeconomic factors such as economic activity, interest rates, and aggregate price levels, most of which are revealed publicly.

The foreign exchange literature implicitly elaborates multiple different interpretations of the private information customers might bring to the market. These vary along three dimensions: (i) whether the information comes from commercial customers, real-money funds, or leveraged investors; (ii) whether the information is fundamental; and (iii) whether the information is passively or actively acquired. Though these three dimensions provide eight conceivable information categories, only some of these appear to be relevant for research. For example, only a small minority of the thousands of non-financial firms around the world would ever attempt to acquire either fundamental or non-fundamental information before trading. The four categories that seem likely to be important, based on the current literature, are discussed below.

Fundamental Information Passively Acquired by Commercial Customers Information about exchange-rate fundamentals may be "dispersed" among customers without being under their control. This hypothesis is most closely associated with Evans and Lyons:

> The dispersed information we have in mind in fact characterizes most variables at the center of exchange rate modeling, such as output, money demand, inflation, [and] consumption preferences … These variables are not realized at the macro level, but rather first as dispersed micro realizations, and only later aggregated by markets and/or governments. For some of these measures, such as risk preferences and money demands, government aggregations of the underlying micro-level shocks do not exist, leaving the full task of aggregation to markets. For other variables, government aggregations exist, but publication lags underlying realizations by 1–4 months, leaving room for market-based aggregation in advance of publication ([61], p. 3).

For concreteness, suppose the economy is expanding rapidly and in consequence commercial firms are all trading actively. Each individual firm might not recognize the generality of its experience but a dealer could potentially see the high economic activity reflected in his commercial-customer order flow. This information would provide the dealer with a signal of GDP concurrent with its realization and thus prior to the associated statistical release.

Fundamental Information Passively Acquired by Financial Customers A variant of the dispersed information hypothesis postulates that the relevant fundamentals concern capital markets as well as the real economy. For example, high demand from institutional investors might indicate that risk aversion is low [58,61,122]. It is not clear whether structural features of financial markets should be considered fundamental, in part because the definition of the term fundamental is not entirely clear. It is clear, however, that any fundamental factor should be relevant to long run equilibrium. Certain structural features of financial markets, like risk appetite, seem likely to influence long-run international macro variables such as international net asset positions (the US net asset position has changed sign but once since 1970), and these in turn seem likely to influence exchange rates. So it seems that some deep financial-market parameters are fundamental, or at least represent some intermediate category between fundamental and non-fundamental.

Fundamental Information Actively Sought by Customers Certain financial customers – typically leveraged investors – forecast exchange rates by combining existing public information with their own economic insights. For example, many such agents attempt to profit from the big returns associated with macro statistical releases by generating private forecasts of upcoming announcements. These customers thus actively generate private fundamental information, rather than passively reflecting information that arises as a normal part of their business. This actively-acquired information could also be reflected in customer order flow, so dealers could still generate their own private signals by observing it. Dealers often report that currency demand is highly correlated within certain types of leveraged investors, permitting them to infer information from observing the trades of just one or a few of these investors.

Indirect evidence for the existence of actively-acquired information comes from Marsh and MacDonald [124]. They find, in a sample of exchange-rate forecasts, that a major cause of forecast heterogeneity "is the idiosyncratic interpretation of widely available information, and that this heterogeneity translates into economically meaningful differences in forecast accuracy" (p. 665). They also find that heterogeneity is a significant determinant of trading volume, consistent with predictions in the literature that diversity of price forecasts generates trading [91,107,181,182].

Non-fundamental Information Some speculative traders may respond to non-fundamental information, like noise traders. Others could respond to non-fundamental hedging needs, as suggested in Bacchetta and van Wincoop [7]. Evidence for the relevance of non-fundamental information is provided in Osler [152], Dominguez and Panthaki [48], and Cao, Evans and Lyons [27]. If the information in order flow is not fundamental it is likely to have only a transitory influence on rates.

Trades based on non-fundamental information may be informative to dealers even if they have only a transitory impact on the market, since dealers speculate at such high frequencies. Indeed, Goodhart [81] insists that dealers rely on nothing but non-fundamental information: dealers' "speculative activities are not based on any consideration of longer-term fundamentals. ... And to repeat, ... the extremely large-scale, very short-term speculative activity in this market by the individual traders ... is *not* based on a long-term future view of economic fundamentals" (pp. 456–457, italics in the original) Consistent with this, US dealers assert that the high-frequency returns on which they focus are unrelated to fundamentals [36]. For example, "at the intraday horizon, PPP has no role according to 93 percent of respondents" (p. 465).

The Evidence: Order Flow Does Carry Information

The evidence indicates fairly clearly that some foreign exchange order flow carries private information. For example, Bjønnes, Osler, and Rime [21] show statistically that banks with the most customer business have an information advantage in the interdealer market, a proposition that dealers themselves certainly support [36,81].

The broader microstructure literature identifies location, specifically proximity to relevant decision-makers, as another potential source of information advantage in financial markets [40,95,129]. Location also appears to be relevant in foreign exchange. Covrig and Melvin [41] find that order flow from Japan tends to lead movements in dollar-yen. Menkhoff and Schmeling [141] find that location affects the information content of interbank trades in the market for rubles. Their analysis indicates that trades originating from the two major financial centers, Moscow and St. Petersburg, have a permanent price impact while trades originating from six peripheral cities do not. D'Souza [49] shows that "trades are most informative when they are initiated in a local country or in major foreign exchange centers of London and New York."

If order flow carries exchange-rate relevant information then one should be able to use it to forecast exchange rates. Studies consistently find that *customer* order flow has predictive power for exchange rates. Evans and Lyons [60] find that daily customer order flow at Citibank has forecasting power for exchange-rate returns at horizons up to one month. Gradojevic and Yang [83] finds that customer and interbank order flow in the Canadian dollar market jointly have forecasting power for exchange rates. They also conclude that a non-linear forecasting structure, specifically an artificial neural network, is superior to linear approaches. Both Evans and Lyons [60] and Gradojevic and Yang [83] conclude that return forecasts are improved when customer order flow is disaggregated according to customer type, which suggests that some participants are more informed than others. Curiously, Rosenberg and Traub [168] provide evidence that *futures* order flow has predictive power for near-term spot returns. This raises the possibility that some informed investors choose to trade in futures markets.

Studies of the forecasting power of *interdealer* order flow arrive at mixed conclusions. Sager and Taylor [170] examine the predictive power of daily interdealer order flow series, including two heavily filtered commercially available order flow series, and the raw interdealer flows examined in Evans and Lyons [58]. They estimate single-equation regressions including order flow and interest differentials as independent variables. Measuring performance in terms of root mean squared error they find that these series do not outperform the random walk when information on future fundamentals is unavailable. In contrast, Rime et al. [166] find that interdealer order flow does outperform the random walk in predicting exchange rates one day ahead. Using three exchange rates (euro-dollar, dollar-yen, sterling-dollar) and associated Reuters (broker) order flow for one year they

Market Microstructure, Foreign Exchange, Table 6

Net purchases for banks in four size categories. The table considers net purchases – the number of purchases minus the number of sales – for four groups of banks vis-à-vis a Scandinavian bank during one week of 1998. Table shows how these net purchases are correlated with contemporaneous returns and with net purchases for other bank categories. All numbers with absolute value over 0.24, 0.28, or 0.36 are significant at the 10 percent, 5 percent, and 1 percent level, respectively. Source: [21]

	Return	Biggest (Rank 1–20)	Big Rank (21–50)	Small (Rank 51–100)	Smallest (Rank > 100)
Return	1.00				
Biggest	0.55***	1.00			
Big	0.26*	0.29**	1.00		
Small	–0.43***	–0.66***	–0.28**	1.00	
Smallest	–0.44***	–0.79***	–0.32***	0.41***	1.00

create forecasts based on what is, in essence, a structural VAR. They use the forecasts to create portfolios of the currencies. For forecast horizons ranging from 14 to 24 hours, the portfolios' Sharpe ratios range from 0.44 to 2.24 and average 1.59. Sharpe ratios for the random walk model and a UIP-based model are generally much lower.

What kind of information is carried by order flow? Evidence is consistent with the presence of both passively-acquired and actively-acquired fundamental information. Evans and Lyons [61] show that Citibank customer order flow has substantial predictive power for US and German GDP growth, inflation, and money growth at horizons ranging up to six months. The results are especially strong at longer horizons, where regressions using only order flow forecast between 21 percent and 58 percent of changes in the fundamental variables. (By contrast, regressions using only the lagged dependent variable or the spot rate generally forecast less than 10 percent.) This suggests that customer order flow concurrently reflects macro fundamentals and that the information may be passively acquired.

Evidence also suggests that order flow carries actively-acquired information about upcoming macro events and news releases. Froot and Ramadorai [74] show that State Street Corporation's institutional-investor flows have significant predictive power for changes in real interest rates at horizons up to thirty days. This would appear to be actively-acquired information.

Rime et al. [166] provide evidence that order flow carries information about upcoming macro news releases. Using thirty different news statistics (fifteen from the US, six from Europe, nine from the UK), the authors run the following regression:

$$\begin{aligned}\mathrm{Ann}^{ki}_{\mathrm{Thurs}+j} &- E_{\mathrm{Thurs}}\mathrm{Ann}^{k}_{\mathrm{Thurs}+j} \\ &= \theta \sum_{i=1}^{j} \mathrm{OrderFlow}_{\mathrm{Thurs}+i} + \psi_{\mathrm{Thurs}+j} \,.\end{aligned}$$

On the left is the news "surprise" for announcement-type k ($k = 1, 2, \ldots, 30$), meaning the difference between the announced figure and the median survey forecast for that announcement. On the right is cumulative interdealer order flow for the period between the survey and the announcement. The estimated relationships are generally quite strong: reported coefficients of determination range up to 0.91 and average 0.45. Since the news releases all lag the realization of the underlying macro aggregate by a month or more, the order flow would not reflect concurrent macro developments but instead appears to have been actively acquired.

This evidence suggests a strong focus on upcoming announcements among speculative agents, a focus that is quite evident in the market. Dealer communication with active customers includes regular – often daily – information on upcoming releases and extensive discussion of the macro context relevant for interpreting these releases. The agents that speculate on such announcements are typically leveraged investors.

Further support for the view that some private information is actively acquired in foreign exchange comes from Osler and Vandrovych [155]. They consider the information in price-contingent orders at the Royal Bank of Scotland with the agents placing those orders disaggregated into eight groups: leveraged investors, institutional investors, large corporations, middle-market corporations, broker-dealers, other banks, the bank's own spot dealers, and the bank's own exotic options desk. The price impact of executed orders, measured as the post-execution return over horizons ranging from five minutes to one week, is evaluated for the three major currency pairs. Results show that orders from leveraged investors have a strong and lasting impact while orders from institutional investors have little or no impact. Consistent with the possible dominance of levered investors, further evidence indicates financial order flow carries more information than commercial order flow, at least at short horizons [28,64,[154].

In short, the evidence is consistent with the hypothesis that customer order flow carries information about macro aggregates that is aggregated by dealers and then reflected in interdealer order flow. The evidence suggests that the customers acquire their information actively and perhaps passively as well.

The Evidence: Is the Information Really Fundamental?

Not all researchers are convinced that the information in foreign exchange order flow is fundamental. Berger et al. [12] highlight their findings (reported earlier) that the long-run price impact of interdealer order flow is smaller than the initial impact, and that explanatory power also declines at longer time horizons. They comment:

> The findings … are consistent with an interpretation of the association between exchange rate returns and order flow as reflecting principally a temporary – although relatively long-lasting – liquidity effect. They are also perhaps consistent with a behavioral interpretation … But our results appear to offer little support to the idea that order flow has a central role in driving long-run fundamental currency values – the 'strong flow-centric' view (p. 9).

Bacchetta and van Wincoop [7] suggest that this interpretation of the result may be more pessimistic than necessary regarding the relevance of fundamental information in order flow. Their model indicates that this pattern would be predicted when order flow reflects both fundamental and non-fundamental information. "In the short run, rational confusion plays an important role in disconnecting the exchange rate from observed fundamentals. Investors do not know whether an increase in the exchange rate is driven by an improvement in average private signals about future fundamentals or an increase in [non-fundamentals]. This implies that [non-fundamentals] have an amplified effect on the exchange rate … " (p. 554)

Evidence presented in Froot and Ramadorai [74] also suggests that the connection from order flow to exchange rates is transitory though long-lasting. Their institutional-flows dataset is large enough to permit a rigorous analysis of order flow and returns at horizons of a year or more (it extends from mid-1994 through early 2001 and covers 18 different currencies vs. the dollar), far longer than horizons considered in most other papers. Like Berger et al. [12], they find that the positive short-run correlation between order flow and returns peaks and then declines. Their correlation estimates reach zero at about 300 trading days and then become statistically negative. The authors note: "[O]ne can interpret the facts as suggesting that any impact of flows on currencies is transitory … [and] any information contained in flows is not about intrinsic value per se (p. 1550)." Since this conclusion is based initially on crude correlations, the authors also undertake a sophisticated VAR decomposition of returns into permanent and transitory components, the results of which lead to the same overall conclusion. This finding cannot be explained in terms of the Bacchetta and van Wincoop [7] insights, since these do not imply the ultimate disappearance of the effect.

Could institutional-investor order flow carry information about macro fundamentals and yet have zero price impact after a year? It was suggested earlier that these observations are consistent when liquidity effects drive the connection from order flow to exchange rates. If real-money funds have roughly a one-year average investment horizon, then the initial upward impact of any, say, purchases – whether or not motivated by fundamental information – would ultimately be offset by a downward impact when the positions are unwound, leaving a zero impact at the one-year horizon. It is also worth noting that Froot and Ramadorai [74] analyze only institutional order flow. As noted earlier, institutional investors typically ignore the currency component of returns when making portfolio allocations, so one would not expect their order flow to have a permanent relation with exchange rates. The trades of other customers might still carry information.

Order flow could also have a transitory influence if exchange-rate expectations are not fully rational, as noted by both Berger et al. [12] and Froot and Ramadorai [74]. A tendency for professional exchange-rate forecasts to be biased and inefficient has been frequently documented [123]. This could explain why exchange rates apparently overreact to certain macro announcements [60]. As in Keynes's beauty contest, short-term traders could profit by correctly anticipating news and how other market participants will react to it, whether or not the reaction to news is rational.

The potential relevance of the behavioral perspective is underscored by extensive evidence for imperfect rationality among currency dealers presented in Oberlechner [147]. Indeed, dealers themselves typically claim that short-run dynamics are driven in part by "excess speculation" [36]. One potential source of excess speculative trading is overconfidence, a human tendency towards which has been extensively documented by psychologists [159]. Odean [150] shows that when agents overestimate the accuracy of their information – a common manifestation of overconfidence – they trade excessively and thereby generate excess volatility. Oberlechner and Osler [148] show, based on a sample of over 400 North American dealers, that currency dealers do not escape the tendency towards overconfidence. Further, they find that overconfident dealers are not driven out of the market: overconfidence is unrelated to a dealer's rank or trading longevity. This suggests that overconfidence may be a permanent structural feature of currency markets.

Information as an Incomplete Explanation

It is important to recognize that "information" is at best a partial explanation for the influence of order flow on exchange rates. An appeal to "information" quickly becomes circular in the absence of a successful economic model of the underlying connections between fundamentals and exchange rates.

This point is best clarified by illustration. Suppose a speculator expects a soon-to-be-released trade balance statistic to be higher than generally expected. According to the information hypothesis, three things happen: (i) the speculator evaluates whether a higher trade balance implies a stronger or weaker home currency and then trades accordingly; (ii) the associated order flow reveals to dealers whether the currency is over- or undervalued; (iii) as more dealers learn the information, it becomes progressively impounded in the exchange rate.

The information research just summarized concentrate on parts (ii) and (iii) of this story. But part (i) is also critical: Speculators must somehow evaluate the implications of the trade balance for the exchange rate in order to choose a position. To accomplish this, the speculator might rely on a model of how fundamentals and exchange rates are connected. But that model cannot itself rely on the information hypothesis without becoming circular: The information hypothesis asserts that exchange rates are determined by order flow because order flow carries information; circularity arises if the information in the order flow is that order flow determines exchange rates, which are determined by information. The speculator might alternatively ignore fundamentals and rely instead on a model of how other people think about fundamentals influence exchange rates. But of course this version of Keynes' beauty contest is equally prone to circularity.

The good news is that models intended to analyze the deep connections between fundamentals and exchange rates can now be based on more than just "assumption and hypotheses" [81]. Instead, they can have well-specified microfoundations based on our new understanding of the structure of currency markets and the exchange-rate determination process. Indeed, in the philosophical outlook of Karl Popper [161], reliance on the best available information is a key test of a model's scientific validity.

Price Discovery in Foreign Exchange

Research so far indicates that order flow influences exchange rates at least in part because it carries information brought to the market by customers. Research has also begun to clarify the exact mechanism through which the information becomes embodied in exchange rates.

Adverse Selection and Customer Spreads

Researchers have tended to assume that the price discovery process in foreign exchange conforms to the process discussed earlier in which adverse selection is key. This view of price discovery has been extensively elaborated in theoretical work, e. g., [100], and many of its predictions are fulfilled in the NYSE [14,89,158].

For structural reasons, this price discovery mechanism cannot apply directly to the foreign exchange market. The mechanism assumes a one-tier market, in which dealers interact only with customers, while foreign exchange is a two-tier market, in which dealers trade with customers in the first tier and trade with each other in the second tier. While this need not imply that adverse selection is entirely irrelevant, it does mean, at a minimum, that the framework needs adjustment before it can be relevant.

Empirical evidence shows that some of the key predictions of adverse selection do not hold in foreign exchange. The framework predicts, for example, that customer spreads are widest for the trades most likely to carry information, which would be large trades and trades with financial customers. The reverse is true, however. Osler et al. [154] analyzes the euro-dollar transactions of a single dealer over four months in 2001 and finds that customer spreads are smaller for large trades and for financial customers. The authors test three other implications of adverse selection, none of which gain support.

Further evidence for an inverse relationship between customer spreads and trade size is provided in Ding [47], which analyzes customer trading on a small electronic communication network. Direct evidence that spreads are narrowest for customer trades that carry the most information comes from Ramadorai [162], which analyzes daily flows through State Street's global custody operations. He finds that asset managers with the greatest skill in predicting (risk-adjusted) returns pay the smallest spreads. Overall it appears that adverse selection does not drive spreads in the customer foreign exchange market.

Adverse selection could, nonetheless, be an important determinant of spreads in the *interdealer* market. Information definitely appears to be asymmetric in that market [21], and the evidence is consistent with the hypothesis that spreads include a significant adverse selection component. Adverse-selection models predict two possible relations between trades and spreads. First, quoted spreads could widen with trade size if trade size is considered informative [52,78,126]. Evidence consistent with this prediction is presented in Lyons [120], but he examined a dealer who exclusively traded in the interdealer market, a form of trading that may no longer exist; later dealer studies fail to confirm this prediction [18,185]. It is possible, however, that trade direction is considered informative even while trade size is not, in which case spreads could still include a significant adverse selection component [99]. This is especially likely in limit-order markets, where the liquidity supplier (limit-order trader) often determines trade size, rather than the liquidity demander (market-order trader). Bjønnes and Rime [18] find strong evidence that trade direction is considered informative in the interdealer market and that adverse selection thereby influences interdealer spreads.

What Drives Customer Spreads?

The apparent irrelevance of adverse selection in the foreign exchange customer market raises an important question: What does drive customer spreads? It appears that structural factors may be at play, since spreads are also widest for the least informed trades in other two-tier markets, including the London Stock Exchange [86], the US corporate bond market [80], and the US municipal bond markets [84,90].

Osler et al. [154] reviews three hypotheses suggested in the broader microstructure literature that could explain this pattern in foreign exchange markets. First, the pattern could reflect the existence of fixed operating costs, which can be covered by a small spread on a large trade or a large spread on a small trade.

Fixed operating costs cannot, however, explain why commercial customers pay higher spreads than financial customers. The "strategic dealing" hypothesis suggests that dealers are strategically subsidizing informed-customer trades in order to gather information they can exploit during later interdealer trading [144,154].

Commercial customers could also pay higher spreads under the "market power" hypothesis of Green et al. [84]. This suggests that dealers have transitory market power relative to customers that do not carefully evaluate their execution quality or who do not know market conditions at the time they trade. Commercial customers in the foreign exchange market tend to be relatively unsophisticated: they are less familiar with standard market practice and typically do not monitor the market on an intraday basis. This may give dealers greater flexibility to extract wider spreads.

Price Discovery in Foreign Exchange

If adverse selection does not describe the price discovery process in foreign exchange, what does? Osler et al. [154] propose an alternative price discovery mechanism consistent with the foreign exchange market's two-tier structure. The mechanism focuses on how dealers choose to offload the inventory accumulated in customer trades. Dealers typically use limit orders to control inventory [18], but not always. Existing theory highlights important determinants of this choice [71,87]: market orders provide speedy execution at the cost of the bid-ask spread, while limit orders provide uncertain execution at an uncertain time but earn the bid-ask spread if execution does take place. This trade-off creates incentives such that market orders are more likely when a dealer's inventory is high, consistent with evidence in Bjønnes and Rime [18] and Osler et al. [154]. It also implies that a dealer should be more likely to place a market order after trading with an informed customer than after trading with an uninformed customer.

To clarify the logic of this second inference, suppose that an informed customer buys from a dealer that previously had zero inventory. That dealer will have three reasons to place a market order in the interdealer market: (i) information that exchange-rate is likely to rise; (ii) a non-zero (and therefore risky) inventory position; and (iii) information that his (short) inventory position is likely to lose value because prices are likely to rise. In consequence, after an informed customer buy transaction the dealer is relatively likely to place a market buy order. This raises the traded price, consistent with the customer's information.

After an uninformed customer purchase, by contrast, a dealer has only one reason to place a market order: risky inventory. If the dealer places a limit order rather than a market order then the uninformed-customer purchase would tend to be associated with negative downward returns, as the limit buy order is executed against a market sell.

One key testable implication of this proposed price discovery mechanism is that the likelihood of an interbank market order is higher after trades that are relatively likely to carry information, specifically financial-customer trades and large trades. Osler et al. [154] finds support for this implication using a probit analysis of their dealer's own trading choices. This indicates that the conditional probability that the dealer places an interbank market order is 9.5 percent for small commercial-customer trades and almost twice as high, at 18.5 percent, after small financial-customer trades. After large financial-customer trades – the most informed of all – the corresponding likelihood is 40.2 percent.

This proposed price discovery mechanism is consistent with much of the empirical evidence discussed so far. For example, it is consistent with the signs of the cointegrating relationships between returns and order flow: positive for financial customers, negative for commercial customers, positive for dealers. The positive cointegration between financial order flow and returns indicates that financial order flow carries fundamental information. The positive cointegration between interdealer order flow and returns suggests that dealers' market orders reflect the information in their customer order flow. The negative cointegration between commercial order flow and returns could also be an outcome of the price discovery hypothesis: if dealers place limit orders after trades with commercial customers (and if commercial customers are indeed relatively uninformed) then a commercial-customer buy will be reflected in an interdealer market sell order, with an associated price decline.

The mechanism is also consistent with Rime et al.'s [166] demonstration that interdealer order flow has strong predictive power for upcoming macro statistical releases, together with other evidence suggesting that leveraged investors bring the most information to the market. If leveraged investors are the most informed customers, then under this price discovery hypothesis interdealer order flow will reflect that group's trades. Since interdealer order flow has strong predictive power for upcoming macro releases, the implication is that leveraged investors devote much effort to forecasting those releases.

Summary and Future Directions

The currency microstructure evidence summarized here provides many new insights about the economics of the currency market and thus the economics of exchange-rate determination. The field thus merits its alternative moniker, "the new microeconomics of exchange rates."

The new evidence reveals that the proximate cause of most exchange-rate dynamics is order flow, which can be interpreted as net liquidity demand. The critical role of order flow is not, of course, in itself an economic explanation for exchange-rate dynamics. Recognizing this, the new literature provides evidence for three economic mechanisms through which order flow could influence exchange rates: inventory effects, liquidity effects, and information.

The information mechanism raises a critical question: What information is carried by order flow? The information apparently originates with customers; dealers then see it reflected in their customer order flow. Some of the information may be dispersed, passively-acquired information about concurrent fundamentals. Some of the information appears to be actively-acquired information about upcoming macro news releases, with the most informative order flow coming from leveraged investors. Some of the information may be non-fundamental.

The literature also investigates the precise mechanism through which a customer's private information becomes reflected in exchange rates. This price discovery mechanism appears to differ strikingly from price discovery on the NYSE, a difference that could reflect a key structural difference across markets: foreign exchange dealers can trade with each other as well as with customers, but the NYSE has no interdealer market.

The literature addresses many questions of importance to researchers in microstructure per se. For example, what determines spreads in foreign exchange? Customer spreads in foreign exchange behave entirely differently from those on, say, the NYSE. On the NYSE, market makers try to protect themselves from informed traders and, if possible, they charge informed traders wider spreads. By contrast, foreign exchange dealers actively court the business of informed traders by quoting them narrow spreads. This could reflect the ability of currency dealers to trade with each other. Currency dealers seek trades with informed customers because the customers' order flow provides information the dealers can exploit in subsequent interdealer trades.

Our knowledge of this market still has big gaps, of course, which provide many fascinating questions for future research. A partial list includes the following:

1. Why do interdealer spreads vary inversely with trading volume and volatility? Does this pattern reflect fixed operating costs, the optimal bunching of liquidity traders, or something else?
2. What determines intraday variations in the price impact of order flow? While it looks like this is strongly influenced by the intraday pattern in interdealer spreads, there is little hard evidence on this point. What other factors might matter?
3. What determines longer-horizon variation in the price impact of order flow? The relevance of this question is enhanced, of course, by the evidence that variation in price impact contributes importantly to the persistence of volatility.
4. There is bound to be substantially more variation across types of financial customers, and across types of corporate customers, than has yet been identified. How

much technical trading is there? What fraction of international investors disregard the currency component of returns when choosing portfolio allocations? Is this fraction changing?

5. There is still much to learn about the nature of the information provided by order flow, how dealers perceive that information, and how dealers use that information. Dealers claim they don't seek and don't use fundamental information but the evidence reveals that much of the information moving through the market is, in fact, related to fundamentals.
6. How strong are inventory, liquidity effects, and information effects in determining the connection between order flow and exchange rates?

Even when these questions have been addressed, however, the larger question – the question that originally motivated foreign exchange microstructure research – will still remain. In dealing with this question the foreign exchange microstructure researchers have followed Karl Popper's [161] agenda for scientific inquiry in its purest form. According to his philosophical perspective, good scientists produce evidence that "falsifies" existing paradigms and then create new paradigms consistent with all the evidence, old and new. The new evidence revealed by currency microstructure has falsified many aspects of traditional macro-based models while shedding new light on the economics of exchange-rate determination.

To develop the next generation of exchange-rate models, researchers now have at their disposal an extensive body of knowledge about how exchange rates are actually determined. This information brings with it the ability – and the responsibility – to construct models with well-specified microfoundations. A rigorous, empirically-relevant paradigm for short-run exchange-rate dynamics is much closer than it was a decade ago.

Bibliography

1. Admati AR, Pfleiderer P (1988) A Theory of Intraday Patterns: Volume and Price Variability. Rev Financial Stud 1:3–40
2. Akram FQ, Rime D, Sarno P (2005) Arbitrage in the Foreign Exchange Markets: Turning on the Microscope. Norges Bank Working Paper 2005-12
3. Andersen TG, Bollerslev T, Francis DX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
4. Anderson TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
5. Artus JR, Knight MD (1984) Issues in the Assessment of the Exchange Rates of Industrial Countries, Occasional Paper 29. International Monetary Fund, Washington, D.C.
6. Austin MP, Bates RG, Dempster MAH, Williams SN (2004) Adaptive Systems for Foreign Exchange Trading. Quant Finance 4:C37–45
7. Bacchetta P, van Wincoop E (2005) Can Information Heterogeneity Explain the Exchange Rate Determination Problem? Am Econ Rev 96:552
8. Baillie R, Bollerslev T (1989) The Message in Daily Exchange Rates: A Conditional Variance Tail. J Bus Econ Stat 7:297–305
9. Bank for International Settlements (2007) Triennial Central Bank Survey of Foreign Exchange and Derivatives Trading Activity. Basle
10. Barberis N, Thaler R (2002) A Survey of Behavioral Finance. NBER Working Paper No. 9222
11. Barker W (2007) The Global Foreign Exchange Market: Growth and Transformation. Bank Can Rev Autumn:3–12
12. Berger D, Chaboud A, Hjalmarsson E, Howorka E (2006) What Drives Volatility Persistence in the Foreign Exchange Market? Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 862
13. Berger D, Chaboud A, Chernenko S, Howorka E, Wright J (2006) Order Flow and Exchange Rate Dynamics in Electronic Brokerage System Data. Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 830
14. Bernhardt D, Hughson E (2002) Intraday trade in Dealership Markets. Euro Econ Rev 46:1697–1732
15. Bertsimas D (1998) Optimal Control of Execution Costs. J Financ Mark 1:1–50
16. Bertsimas D, Andrew WL (1998) Optimal Control of Execution Costs. J Financial Markets 1:1–50
17. Bertsimas, Lo A (1997)
18. Bjønnes GH, Rime D (2005) Dealer Behavior and Trading Systems in Foreign Exchange Markets. J Financial Econ 75:571–605
19. Bjønnes GH, Rime D, Solheim HOA (2005) Liquidity Provision in the Overnight Foreign Exchange Market. J Int Money Finance 24:175–196
20. Bjønnes GH, Rime D, Solheim HOA (2005) Volume and Volatility in the FOREIGN EXCHANGE Market: Does it Matter Who You Are? In: De Grauwe P (ed) Exchange Rate Modeling: Where Do We Stand? MIT Press, Cambridge
21. Bjønnes G, Osler C, Rime D (2007) Asymmetric Information in the Interdealer Foreign Exchange Market. Presented at the Third Annual Conference on Market Microstructure, Budapest, Hungary, 15 Sept 2007
22. Black, Fischer (1986) Noise. Finance 41:529–543
23. Bollerslev T, Melvin M (1994) Bid-ask Spreads and Volatility in the Foreign Exchange Market. J Int Econ 36:355–372
24. Brandt M, Kavajecz K (2005) Price Discovery in the US Treasury Market: The Impact of Order Flow and Liquidity on the Yield Curve. J Finance 59:2623–2654
25. Cai J, Cheung YL, Raymond SKL, Melvin M (2001) Once-in-a Generation Yen Volatility in 1998: Fundamentals, Intervention, and Order Flow. J Int Money Finance 20:327–347
26. Campbell JY, Shiller RJ (1988) Stock Prices, Earnings, and Expected Dividends. J Finance 43:661–676
27. Cao H, Evans M, Lyons KR (2006) Inventory Information. J Bus 79:325–363
28. Carpenter A, Wang J (2003) Sources of Private Information in FX Trading. Mimeo, University of New South Wales
29. Carlson JA, Melody L (2006) One Minute in the Life of the

DM/US$: Public News in an Electronic Market. J Int Money Finance 25:1090–1102

30. Carlson JA, Dahl C, Osler C (2008) Short-Run Exchange-Rate Dynamics: Theory and Evidence. Typescript, Brandeis Univ
31. Chaboud AP, Chernenko SV, Howorka E, Iyer KRS, Liu D, Wright JH (2004) The High-Frequency Effects of US Macroeconomic Data Releases on Prices and Trading Activity in the Global Interdealer Foreign Exchange Market. Federal Reserve System Board of Governors, International Finance Discussion Papers Number 823
32. Chakravarty S (2000) Stealth Trading: Which Traders' Trades Move Stock Prices? J Financial Econ 61:289–307
33. Chang Y, Taylor SJ (2003) Information Arrivals and Intraday Exchange Rate Volatility. Int Financial Mark Inst Money 13:85–112
34. Chaboud A, Weinberg S (2002) Foreign Exchange Markets in the 1990s: Intraday Market Volatility and the Growth of Electronic Trading. B.I.S. Papers No. 12
35. Chaunzwa MJ (2006) Investigating the Economic Value Added of More Advanced Technical Indicators. Typescript, Brandeis University
36. Cheung YW, Chinn MD (2001) Currency Traders and Exchange Rate Dynamics: A Survey of the US Market. J Int Money Finance 20:439–471
37. Cheung YW, Chinn MD, Marsh I (2004) How Do UK-based Foreign Exchange Dealers Think Their Market Operates? Int J Finance Econ 9:289–306
38. Chordia T, Roll R, Subrahmanyam A (2002) Order Imbalance, Liquidity, and Market Returns. J Financial Econ 65:111–130
39. Copeland T, Galai D (1983) Information Effects on the Bid-Ask Spread. J Finance 38:1457–1469
40. Coval JD, Moskowitz TJ (2001), The Geography of Investment: Informed Trading and Asset Prices. J Political Econ 109(4):811–841
41. Covrig V, Melvin M (2002) Asymmetric Information and Price Discovery in the FX Market: Does Tokyo Know More About the Yen? J Empir Financ 9:271–285
42. Covrig V, Melvin M (2005) Tokyo Insiders and the Informational Efficiency of the Yen/Dollar Exchange Rate. Int J Finance Econ 10:185–193
43. Cross S (1998) All About … The Foreign Exchange Market in the United States. Federal Reserve Bank of New York. http://www.newyorkfed.org/education/addpub/usfxm/
44. Daniélsson J, Love R (2006) Feedback Trading. Int J Finance Econ 11:35–53
45. Delong B, Shleifer A, Summers L, Waldmann RJ (1990) Positive Feedback Investment Strategies and Destabilizing Rational Speculation. J Finance 45:379–395
46. Dewachter H (2001) Can Markov Switching Models Replicate Chartist Profits in the Foreign exchange Market? J Int Money Financ 20:25–41
47. Ding L (2006) Market Structure and Dealer's Quoting Behavior in the Foreign Exchange Market. Typescript, University of North Carolina at Chapel Hill
48. Dominguez K, Panthaki F (2006) What Defines News in Foreign Exchange Markets? J Int Money Finance 25:168–198
49. D'Souza C (2007) Where Does Price Discovery Occur in FX Markets? Bank of Canada Working Paper 2007-52
50. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Raturns? J Bank Finance 31:279–296
51. Dunne P, Hau H, Moore M (2007) A Tale of Two Plattforms: Interdealer and Retail Quotes in the European Bond Markets. Presented at the Third Annual Conference on Microstructure, Magyar Bank, Budapest, September 2007
52. Easley D, O'Hara M (1987) Price Trade Size, and Information in Securities Markets. J Financial Econ 19:69–90
53. Ederington L, Jae HL (2001) Intraday Volatility in Interest-Rate and Foreign-Exchange Markets: ARCH, Announcement, and Seasonality Effects. J Futures Mark 21:517–552
54. Ehrmann M, Fratzscher M (2005) Exchange Rates and Fundamentals: New Evidence from Real-Time Data. J Int Money Finance 24:317–341
55. Euromoney (2006) FX Poll. http://www.euromoney.com/article.asp?ArticleID=1039514
56. Evans M (2002) FX Trading and Exchange Rate Dynamics. J Finance 57:2405–2448
57. Evans M, Lyons RK (2002) Information Integration and FX Trading. J Int Money Financ 21:807–831
58. Evans M, Lyons KR (2002) Order Flow and Exchange Rate Dynamics. J Political Econ 110(2002):170–180
59. Evans M, Lyons KR (2005) Do Currency Markets Absorb News Quickly? J Int Money Financ 24:197–217
60. Evans M, Lyons KR (2005) Meese-Rogoff Redux: Micro-Based Exchange-Rate Forecasting. Am Econ Rev Papers Proc 95:405–414
61. Evans M, Lyons KR (2007) Exchange-Rate Fundamentals and Order Flow. NBER Working Paper 13151
62. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? J Financ Econ, forthcoming
63. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? forthcoming, Journal of Financial Econmics
64. Fan M, Lyons RK (2003) Customer Trades and Extreme Events in Foreign exchange. In: Paul Mizen (ed) Monetary History, Exchange Rates and Financial Markets: Essays in Honour of Charles Goodhart. Edward Elgar, Northampton, pp 160–179
65. Federal Reserve Bank of New York (2004) Managing Operational Risk in Foreign Exchange
66. Federal Reserve Bank of New York (2007) http://www.newyorkfed.org/xml/gsds_transactions.html
67. Fleming M (2003) Measuring Treasury Market Liquidity. Federal Reserve Bank of New York. Econ Policy Rev 9:83–108
68. Fleming M, Mizrach B (2007) The Microstructure of a UST-reasury ECN: The BrokerTec Platform. Typescript, Federal Reserve Bank of New York
69. Flood RP, Taylor MP (1996) Exchange-Rate Economics: What's Wrong with the Conventional Macro Approach. In: Jeffrey A Frankel, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp. 261–301
70. Foster DF, Viswanathan S (1993) Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models. J Finance 3:187–211
71. Foucault T (1999) Order Flow Composition and Trading Costs in a Dynamic Limit Order Market. J Financial Mark 2:99–134
72. Frankel JA, Galli G, Giovannini A (1996) Introduction. In: Frankel JA, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago pp. 1–15
73. Frenkel JA, Mussa ML (1985) Asset markets, exchange rates, and the balance of payments. NBER working paper 1287

74. Froot K, Ramadorai Tarun (2005) Currency Returns, Intrinsic Value, and Institutional-Investor Flows. Finance 55:1535–1566
75. Galati G (2000) Trading Volume, Volatility, and Spreads in Foreign Exchange Markets: Evidence from Emerging Market Countries. BIS Working Paper 93
76. Gau YF (2005) Intraday Volatility in the Taipei Foreign Exchange Market. Pac Basin Finance J 13:471–487
77. Gehrig T, Menkhoff L (2004) The Use of Flow Analysis in Foreign Exchange: Exploratory Evidence. J Int Money Finance 23:573–594
78. Glosten L (1989) Insider Trading, Liquidity, and the Role of the Monopolist Specialist. J Bus 62:211–235
79. Glosten LR, Milgrom PR (1985) Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. J Financial Econ 14:71–100
80. Goldstein MA, Hotchkiss ES, Sirri ER (2007) Transparency and Liquidity: A Controlled Experiment on Corporate Bonds. Rev Financial Stud 20:235–273
81. Goodhart C (1988) The Foreign Exchange Market: A Random Walk with a Dragging Anchor. Economica 55:437–60
82. Goodhart CAE, Hall SG, Henry SGB, Pesaran B (1993) News Effects in High-Frequency Model of the Sterling-Dollar Exchange Rate. J Appl Econ 8:1–13
83. Gradojevic N, Yang J (2006) Non-Linear, Non-Paramettric, Non-Fundamental Exchange Rate Forecasting. J Forecast 25:227–245
84. Green RC, Hollifield B, Schurhoff N (2007) Financial Intermediation and the Costs of Trading in an Opaque Market. Rev Financial Stud 20:275–314
85. Hansch O, Naik N, Viswanathan S (1998) Do Inventories Matter in Dealership Markets? Some Evidence from the London Stock Exchange. J Finance 53:1623–1656
86. Hansch O, Naik N, Viswanathan S (1999) Preferencing, Internalization, Best Execution, and Dealer Profits. J Finance 54:1799–1828
87. Harris L (1998) Optimal Dynamic Order Submission Strategies in Some Stylized Trading Problems. Financial Mark Inst Instrum 7:1–75
88. Harris L, Gurel E (1986) Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures. J Finance 41:815–29
89. Harris L, Hasbrouck J (1996) Market vs. Limit orders: The SuperDOT Evidence on Order Submission Strategy. J Financial Quant Anal 31:213–231
90. Harris LE, Piwowar MS (2006) Secondary Trading Costs in the Municipal Bond Market. J Finance 61:1361–1397
91. Harris M, Raviv A (1993) Differences of Opinion Make a Horse Race. Rev Financial Stud 6:473–506
92. Hartmann P (1999) Trading Volumes and Transaction Costs in the Foreign Exchange Market: Evidence from Daily Dollar-Yen Spot Data. J Bank Finance 23:801–824
93. Hasbrouck J (1991) Measuring the Information Content of Stock Trades. J Finance 46:179–220
94. Hashimoto Y (2005) The Impact of the Japanes Banking Crisis on the Intraday FOREIGN EXCHANGE Market in Late 1997. J Asian Econ 16:205–222
95. Hau H (2001) Location Matters: An Examination of Trading Profits. J Finance 56(3):1959–1983
96. Hau H, Rey H (2003) Exchange Rates, Equity Prices, and Capital Flows. Rev Financial Stud 19:273–317
97. Hau H, Killeen W, Moore M (2002) How has the Euro Changed the Foreign Exchange Market? Econ Policy, issue 34, pp 151–177
98. Hendershott T, Seasholes M (2006) Market Maker Inventories and Stock Prices. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
99. Huang RD, Stoll HR (1997) The Components of the Bid-Ask Spread: A General Approach. Rev Financial Stud 10:995–1034
100. Holden CW, Subrahmanyam A (1992) Long-Lived Private Information and Imperfect Competition. J Finance 47:247–270
101. Holthausen RW, Leftwich RW, Mayers D (1990) Large-Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects. J Financial Econ 26:71–95
102. Ito T (1990) Foreign Exchange Rate Expectations: Micro Survey Data. Am Econ Rev 80:434–449
103. Ito T, Hashimoto Y (2006) Intraday Seasonality in Activities of the Foreign Exchange Markets: Evidence from the Electronic Broking System. J Jap Int Econ 20:637–664
104. Jones CM, Kaul G, Lipson ML (1994) Transactions, Volume and Volatility. Rev Financial Stud 7:631–651
105. Jorion P (1996) Risk and Turnover in the Foreign Exchange Market. In: Frankel JA, Galli G, Giovaninni A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp 19–37
106. Jovanovic B, Rousseau PL (2001) Liquidity Effects in the Bond Market. Federal Reserve Bank of Chicago Econ Perspect 25:17–35
107. Kandel E, Pearson ND (1995) Differential Interpretation of Public Signals and Trade in Speculative Markets. J Political Econ 103:831–872
108. Kearns J, Manners P (2006) The Impact of Monetary Policy on the Exchange Rate: A Study Using Intraday Data. International Journal of Central Banking 2:175–183
109. Killeen W, Lyons RK, Moore M (2005) Fixed versus Flexible: Lessons from EMS Order Flow. Forthcoming, J Int Money Finance
110. Killeen W, Lyons RK, Moore M (2006) Fixed versus Flexible: Lessons from EMS Order Flow. J Int Money Financ. 25:551–579
111. Kim M, Kim M (2001) Implied Volatility Dynamics in the Foreign Exchange Markets. J Int Money Finance 22:511–528
112. Klitgaard T, Weir L (2004) Exchange Rate Changes and Net positions of Speculators in the Futures Market. Federal Reserve Bank of New York Econ Policy Rev 10:17–28
113. Kothari SP (2001) Capital Markets Research in Accounting. J Account Econ 31:105–31
114. Kuhn TS (1970) The Structure of Scientific Revolutions, 2nd. edn. University of Chicago Press, Chicago
115. Kyle A (1985) Continuous Auctions and Insider Trading. Econometrica 53:1315–1335
116. Lane PR (2001) The New Open Economy Macroeconomics: A Survey. J Int Econ 54:235–266
117. LeBaron B (1998) Technical Trading Rules and Regime Shifts in Foreign Exchange. In Acar E, Satchell S (eds) Advanced Trading Rules. Butterworth-Heinemann, pp. 5–40
118. Love R, Payne R (2003) Macroeconomic News, Order Flows, and Exchange Rates. Typescript, London School of Economics
119. Lui Yu-Hon, Mole D (1998) The Use of Fundamental and Technical Analyses by Foreign Exchange Dealers: Hong Kong Evidence. J Int Money Financ 17:535–45

120. Lyons RK (1995) Tests of Microstructural Hypotheses in the Foreign Exchange Market. J Financ Econ 39:321–351
121. Lyons RK (1997) A Simultaneous Trade Model of the Foreign Exchange Hot Potato. J Int Econ 42:275–98
122. Lyons RK (2001) The Microstructure Approach to Exchange Rates. MIT Press, Cambridge and London
123. MacDonald R (2000) Expectations Formation and Risk in Three Financial Markets: Surveying What the Surveys Say. J Econ Surv 14:69–100
124. MacDonald R, Marsh I (1996) Currency Forecasters are Heterogeneous: Confirmation and Consequences. J Int Money Finance 15:665–685
125. Madhavan AN, Cheng M (1997) In Search of Liquidity: Block Trades in the Upstairs and Downstairs Markets. Rev Financial Stud 10:175–203
126. Madhavan AN, Smidt S (1991) A Bayesian Model of Intraday Specialist Pricing. J Financial Econ 30:99–134
127. Madhavan AN, Smidt S (1993) An Analysis of Changes in Specialist Inventories and Quotations. J Finance 48:1595–1628
128. Madhavan A, Richardson M, Roomans M (1997) Why Do Security Prices change? A Transaction-Level Analysis of NYSE Stocks. Rev Financial Stud 10:1035–1064
129. Malloy C (2005) The Geography of Equity Analysis. J Finance 60(2):719–756
130. Manaster S, Mann SC (1996) Life in the Pits: Competitive Market Making and Inventory Control. Rev Financial Stud 9:953–975
131. Marsh I, O'Rourke C (2005) Customer Order Flow and Exchange Rate Movements: Is There Really Information Content? Presented at the Norges Bank Conference on Equity and Currency Microstructure, Oslo
132. Martens M (2001) Forecasting Daily Exchange Rate Volatility Using Intraday Returns. J Int Money Financ 20:1–23
133. Fratscher M (2007) US Shocks and Global Exchange Rate Configurations. Presented at the NBERIFM meetings, Cambridge, MA, 10 July 2007
134. Meese RA, Rogoff K (1983) Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample? J Int Econ 14:3–24
135. Mende A (2006) 09/11 and the USD/EUR Exchange Market. Appl Financial Econ 16:213–222
136. Mende A, Menkhoff L (2006) Profits and Speculation in Intra-Day Foreign Exchange Trading. J Financial Mark 9:223–245
137. Menkhoff L (1997) Examining the Use of Technical Currency Analysis. Int J Finance Econ 2:307–318
138. Menkhoff L (2001) Importance of Technical and Fundamental Analysis in Foreign Exchange Markets. Int J Finance Econ 6:81–93
139. Menkhoff L, Gehrig T (2006) Extended Evidence on the Use of Technical Analysis in Foreign Exchange. Int J Finance Econ 11:327–38
140. Menkhoff L, Taylor MP (2006) The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. University of Warwick, Department of Economics, The Warwick Economics Research Paper Series (TWERPS)
141. Menkhoff L, Osler C, Schmeling M (2007) Order-Choice Dynamics under Asymmetric Information: An Empirical Analysis. Typescript, Brandeis University
142. Milgrom P, Stokey N (1982) Information, Trade, and Common Knowledge. J Econ Theory 26:17–27
143. Morris S (1984) Trade with Heterogeneous Prior Beliefs and Asymmetric Information. Econometrica 62:1327–1347
144. Naik NY, Neuberger A, Viswanathan S (1999) Trade Disclosure Regulation in Markets With Negotiated Trades. Rev Financial Stud 12:873–900
145. New York Stock Exchange (2007) Historical Facts and Statistics. www.nysedata.com/nysedata/Default.aspx?tabid=115
146. Oberlechner T (2001) Evaluation of Currencies in the Foreign Exchange Market: Attitudes and Expectations of Foreign Exchange Traders. Z Sozialpsychologie 3:180–188
147. Oberlechner T (2004) The Psychology of the Foreign Exchange Market. Wiley, Chichester
148. Oberlechner T, Osler C (2007) Overconfidence in Currency Markets. Typescript, Brandeis International Business School
149. Obstfeld M, Rogoff K (1995) Exchange Rate Dynamics Redux. J Political Econ 3:624–660
150. Odean T (1998) Volume, Volatility, Price, and Profit: When All Traders Are Above Average. J Finance 53:1887–1934
151. Osler CL (2003) Currency Orders and Exchange-Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. J Finance 58:1791–1819
152. Osler CL (2005) Stop-Loss Orders and Price Cascades in Currency Markets. J Int Money Finance 24:219–41
153. Osler CL (2006) Macro Lessons from Microstructure. Int J Finance Econ 11:55–80
154. Osler CL, Savaser T (2007) The Microstructure of Extreme Exchange-Rate Returns. Typescript, Brandeis International Business School
155. Osler CL, Vandrovych V (2007) Which Customers Bring Information to the in Foreign Exchange Market? Typescript, Brandeis International Business School
156. Pasquariello P, Vega C (2005) Informed and Strategic Order Flow in the Bond Markets, Working Paper
157. Payne R (2003) Informed trade in Spot Foreign Exchange Markets: An Empirical Investigation. J Int Econ 61:307–329
158. Peterson MA, Sirri ER (2003) Order Preferencing and Market Quality on US Equity Exchanges. Rev Financial Stud 16:385–415
159. Plous S (1993) The psychology of judgment and decision making. McGraw-Hill, New York
160. Pong S, Shackleton MB, Taylor SJ, Xu X (2004) Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models. J Bank Financ 28:2541–2563
161. Popper K (1959) The Logic of Scientific Discovery. Routledge, United Kingdom
162. Ramadorai T (2006) Persistence, performance, and prices in foreign exchange markets. Oxford University Working Paper
163. Reiss PC, Werner IM (1995) Transaction Costs in Dealer Markets: Evidence from the London Stock Exchange. In: Andrew L (ed) The Industrial Organization and Regulation of the Securities Industry. University of Chicago Press, Chicago
164. Reiss PC, Werner IM (1998) Does risk sharing motivate interdealer trading? J Finance 53:1657–1703
165. Reiss PC, Werner IM (2004) Anonymity, adverse selection, and the sorting of interdealer trades. Rev Financial Stud 18:599–636
166. Rime D, Sarno L, Sojli E (2007) Exchange-rate Forecasting, Order Flow, and Macro Information. Norges Bank Working Paper 2007-2
167. Rogoff KS (1996) The Purchasing Power Parity Puzzle. J Econ Lit 34:647–668

168. Rosenbert JV, Traub LG (2006) Price Discovery in the Foreign Currency Futures and Spot Market. Typescript, Federal Reserve Bank of New York
169. Sager M, Taylor MP (2006) Under the Microscope: The Structure of the Foreign Exchange Market. Int J Finance Econ 11:81–95
170. Sager M, Taylor MP (2008) Commercially Available Order Flow Data and Exchange Rate Movements: Caveat Emptor. J Money Credit Bank 40:583–625
171. Savaser T (2006) Exchange Rate Response to Macro News: Through the Lens of Microstructure. Presented at the Bank of Canada Workshop on Equity and Currency Microstructure, October 2006
172. Savaser T (2007) Stop-Loss Orders and Macro News in Currency Markets. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
173. Shiller R (1981) Do Stock Prices Move Too Much to Be Jusdtified by Subseuqent Changes in Dividends? Am Econ Rev 71:421–435
174. Shleifer A (1986) Do Demand Curves Slope Down? J Finance 41:579–90
175. Simon DP (1991) Segmentation in the Treasury Bill Market: Evidence from Cash Management Bills. J Financial Quant Anal 26:97–108
176. Simon DP (1994) Further Evidence on Segmentation in the Treasury Bill Market. J Bank Finance 18:139–151
177. Stoll H (1978) The Supply of Dealer Services in Securities Markets. J Finance 33:1133–1151
178. Taylor A, Farstrup A (2006) Active Currency Management: Arguments, Considerations, and Performance for Institutional Investors. CRA Rogers Casey International Equity Research, Darien Connecticut
179. Taylor MP (2002) Purchasing Power Parity and the Real Exchange Rate. IMF Staff Papers 48:65–105
180. Taylor MP, Allen H (1992) The Use of Technical Analysis in the Foreign Exchange Market. J Int Money Finance 11:304–314
181. Varian HR (1985) Divergence of Opinion in Complete Markets: A Note. J Finance 40:309–317
182. Varian HR (1989) Differences of Opinion in Financial Markets. In: Stone CC (ed) Financial Risk: Theory, Evidence, and Implications. Federal Reserve Bank of St. Louis, pp. 3–37
183. Wudunn S (1995) Japanese Delayed Letting US Know of Big Bank Loss. New York Times, 10 Oct
184. Yan B, Zivot E (2007) The Dynamics of Price Discovery. Typescript, University of Washington
185. Yao JM (1998) Market making in the interbank foreign exchange market. Stern School of Business, New York University Working Paper S-98

Microeconometrics

Pravin K. Trivedi
Department of Economics, Indiana University,
Bloomington, USA

Article Outline

Glossary

Cowles commission approach An approach to structural econometric modeling identified with the pioneering work of the Cowles Foundation during the 1940s and 1950s.

Endogenous variable A variable whose value is determined within a specified model.

Exogenous A variable that is assumed given for the purposes of analysis because its value is determined outside the model of interest.

Reduced form models A stochastic model with relationships between endogenous variables on the one hand and all exogenous variables on the other.

Structural model A stochastic model with interdependent endogenous and exogenous variables.

Treatment effects An effect attributed to a change in the value of some policy variable analogous to a treatment in a clinical trial.

Definition of the Subject

Microeconometrics deals with model-based analysis of individual-level or grouped data on the economic behavior of individuals, households, establishments or firms. Regression methods applied to cross-section or panel (longitudinal) data constitute the core subject matter. Microeconometric methods are also broadly applicable to social and mathematical sciences that use statistical modeling. The data used in microeconometric modeling usually come from cross section and panel surveys, censuses, or social experiments. A major goal of microeconometric analysis is to inform matters of public policy. The methods of microeconometrics have also proved useful in providing model-based data summaries and prediction of hypothetical outcomes.

Introduction

Microeconometrics takes as its subject matter the regression-based modeling of economic relationships using data at the levels of individuals, households, and firms. A distinctive feature microeconometrics derives from the low level of aggregation in the data. This has immediate implications for the functional forms used to model analyze the relationships of interest. Disaggregation of data brings to the forefront heterogeneity of individuals, firms, and organizations. Modeling such heterogeneity is often essential for making valid inferences about the underlying relationships. Typically aggregation reduces noise and leads to smoothing due to averaging of movements in opposite directions whereas disaggregation leads to loss of continuity and smoothness. The range of variation in micro data is also typically greater. For example, household's average weekly consumption of (say) meat is likely to vary smoothly, while that of an individual household in a given week may be frequently zero, and may also switch to positive values from time to time. Thus, micro data exhibit "holes, kinks and corners" [80]. The holes correspond to nonparticipation in the activity of interest, kinks correspond to the switching behavior, and corners correspond to the incidence of nonconsumption or nonparticipation at specific points of time. Consequently, discreteness and nonlinearity of response are intrinsic to microeconometrics.

Another distinctive feature of microeconometrics derives from the close integration of data and statistical modeling assumptions employed in analyzing them. Sample survey data, the raw material of microeconometrics, are subject to problems of complex survey methodology, departures from simple random sampling assumptions, and problems of sample selection, measurement errors, incomplete and/or missing data – problems that in principle impede the generalization from sample to population. Handling such issues is an essential component of microeconometric methodology.

An important application of microeconometrics is to tests predictions of microeconomic theory. Tests based on micro data are more attractive and relatively more persuasive because (a) the variables involved in such hypotheses can be measured more directly, (b) the hypotheses under test are likely to be developed from theories of individual behavior, and (c) a realistic portrayal of economic activity should accommodate a broad range of outcomes and responses that are a consequence of individual heterogeneity and that are predicted by underlying theory. In many public policy issues one is interested in the behavioral responses of a specific group of economic agents under some specified economic environment. One example is the impact of unemployment insurance on the job search behavior of young unemployed persons. To address these issues directly it is essential to use micro data.

The remainder of this article is organized as follows. In the next section I provide a historical perspective of the development of microeconometrics and sketch the topics in which important advances have occurred. In Sect. "Historical Background" we detail two models – the discrete choice model and the selection model – that are landmark developments in microeconometrics and provide important reference points for the remainder of the article. Sect. "Two Leading Examples" outlines three dominant modeling methodologies for structural modeling in microeconometrics. The final Sect. "Causal Modeling" surveys some of the major challenges in microeconometrics and the available modeling tools for dealing with these challenges. To stay within space constraints, I emphasize developments that have influenced microeconomic data analysis, and pay less attention to general theoretical analyzes.

Historical Background

Analysis of individual data has a long history. Engel [23], Allen and Bowley [2], Houthakker [43], and Prais and Houthakker [79] all made pioneering contributions to the research on consumer behavior using household budget data. Other seminal studies include Marschak and Andrews [77] in production theory, and Stone [86], and Tobin [88] in consumer demand. Nevertheless, the pathbreaking econometric developments initiated by the Cowles Foundation during the 1940s and 1950s were motivated by concerns of macroeconomic modeling. The initial impact of this research was therefore largely on the development of large-scale multi-equation aggregate models of sectors and the economy. Although the Cowles Commission work was centered on the linear simultaneous equations model (SEM), while modern microeconometrics emphasizes nonlineairties and discreteness, the SEM conceptual framework has proved to be a crucial and formative influence in structural microeconometric modeling.

The early microeconometric work, with the important exception of Tobin [88], relied mainly on linear models, with little accommodation of discreteness, kinks, and corners. Daniel McFadden's [68] work on analysis of discrete choice and James Heckman's [30,31,32,33] work on models of truncation, censoring and sample selection, which combined discrete and continuous outcomes, were pathbreaking developments that pioneered the development of modern microeconometrics. These developments overlapped with the availability of large micro data sets beginning in the 1950s and 1960s.

These works were a major departure from the overwhelming reliance on linear models that characterized earlier work. Subsequently, they have led to major methodological innovations in econometrics. Among the earlier textbook level treatment of this material (and more) are Maddala [76] and Amemiya [3]. As emphasized by Heckman [35], McFadden [73] and others, many of the fundamental issues that dominated earlier work based on market data remain important, especially concerning the conditions necessary for identifiability of causal economic relations. But the style of microeconometrics is sufficiently distinct to justify writing a text that is exclusively devoted to it.

Modern microeconometrics based on individual, household, and establishment level data owes a great deal to the greater availability of data from cross section and longitudinal sample surveys and census data. In the last two decades, with the expansion of electronic recording and collection of data at the individual level, data volume has grown explosively. So too has the available computing power for analyzing large and complex data sets. In many cases event level data are available; for example, marketing science often deals with purchase data collected by electronic scanners in supermarkets, and industrial organization literature contains econometric analyzes of airline travel data collected by online booking systems. New branches of economics, such as social experimentation and experimental economics, have opened up that generate "experimental" data. These developments have created many new modeling opportunities that are absent when only aggregated market level data are available. At the same time the explosive growth in the volume and types of data has also given rise to numerous methodological issues. Processing and econometric analysis of such large micro data bases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeconometrics. Econometric analysis of such data is the subject matter of this book.

Areas of Advances

Both historically and currently, microeconometrics concentrates on the so-called limited dependent variable (LDV) models. The LDV class deals with models in which the outcome of interest has a limited range of variation, in contrast to the case where variation is continuous and on the entire real line. Examples are binary valued outcomes, polychotomous outcomes, non-negative integer-valued outcomes, and truncated or censored variables where values outside a certain range are not observed. An example of censoring arises in modeling the labor supply of working women. Here the data refers to the number of hours of work of the employed women even though from an empirical perspective the economist is interested in both the decision to participate in the labor force (extensive margin) and also in the choice of hours of work (intensive margin) conditional on participation. From this perspective the sample on hours of work is censored and the analysis of hours of work of only those who participate potentially suffers from "selection bias". Analysis of transitions between states and of time spent in a state, e. g. unemployment, using the methods of hazard (survival) analysis also confronted the issue of truncation and censoring, since in many cases the spells of unemployment (durations) were only partially observed. Many economic outcomes such as choice of occupation or travel model, and event counts are inherently discrete and hence fall in the LDV class. Many others involve interdependent discrete and continuous outcomes, e. g. participation and hours of work.

- LDV topics have maintained their core status in the area. But their scope has expanded to include count data models [12] and a much wider variety of selection models. Whereas in 1975 virtually all of the models of discrete choice were static and cross sectional, now discrete choice analysis has developed in many directions, including dynamic aspects which permit dependence between past, current and future discrete choices. Dynamic discrete choice modeling is now embedded in dynamic programming models [22,83]. Individuals often state their preferences over hypothetical choices (as when they are asked to reveal preferences over goods and services not yet in the market place), and they also reveal their preferences in the market place. Modern discrete choice analysis integrates stated preferences and revealed preferences [89,90].
- In 1975 the subject of multivariate and structural estimation of discrete response models required further work in almost every respect. In modern microeconometric models generally, and discrete choice models specifically, there is greater emphasis on modeling data using flexible functional forms and allowing for heterogeneity. This often leads to mixture versions of these models. Advances in computer hardware and software technologies have made simulation-based methods of all types, including Bayesian Markov chain Monte Carlo methods, more accessible to practitioners. Varieties of LDV models that were previously outside the reach of practitioners are now widely used. Inference based on resampling methods such as bootstrap that do not require closed form expressions for asymptotic variances are now quite common in microeconometrics.
- Extensions of many, if not most, LDV models to allow for panel data are now available [44]. Random effects panel models are especially amenable to simulation-based estimation. There have been important advances in handling advanced linear panel data models (including dynamic panels) and nonlinear panel data models – especially models for binary and multinomial outcomes, censored variables, count variables, all of which are now more accessible to practitioners.
- Bayesian approaches are well-suited for analyzing complex LDV because they efficiently exploit the underlying latent variable structure. Bayesian analysis of LDV models is well-developed in the literature, but its incorporation into mainstream texts still lags [53]. Specialized monographs and texts, however, fill this gap.
- Treatment evaluation, which deals with measurement of policy impact at micro level, is now conspicuous and major new topic. The impact of the topic is broad because treatment evaluation is discussed in the context of many different LDV models, using a variety of parametric and semi- or nonparametric approaches, under a variety of different assumptions about the impact of treatment. The literature on this topic is now very extensive, see Heckman and Robb [36], Imbens and Angrist [45], Heckman and Vytlacil [39], and Lee [58] for a monograph-length treatment.
- Topic related to data structures now receive more attention. This includes the pros and cons of observational data and those from social and natural experiments. These topics arise naturally in the context of treatment evaluation. Other data related topics such as survey design and methodology, cross sectional and spatial dependence, clustered observations, and missing data also get greater attention.
- As regards estimation and inference, the classical methods of maximum likelihood, least squares and method of moments were previously dominant, with some exceptions. These methods typically make strong distributional and functional form assumptions that are often viewed with skepticism because of their potential impact on policy conclusions. By contrast, there is now a greater variety of semiparametric estimators in use, of which quantile regression is a leading example [51].

Nonparametric regression is another new topic. There is now a large literature dealing with most standard models and issues from a semi-parametric viewpoint.

Two Leading Examples

To illustrate some salient features of microeconometrics, the structure of two leading models, the first one for discrete choice and the second for sample selection, will be described and explained. Latent variables play a key role in the specification of both models, and in the specification of LDV models more generally. Distributional and structural restrictions are usually imposed through the latent variable specifications. Estimation of the models can also exploit the latent variable structure of such models.

Example 1: Random Utility Model

McFadden played a major role in the development of the random utility model (RUM) that provides the basis of discrete choice analysis; see McFadden [68,70,71,72]. Discrete choice models, firmly established in the analysis of transport mode choice, are now used extensively to model choice of occupations, purchase of consumer durables and brand choice.

The RUM framework is an extension of Thurstone [87]. In the binary RUM framework the agent chooses between alternatives 0 and 1 according to which leads to higher satisfaction or utility which is treated as a latent variable. The observed discrete variable y then takes value 1 if alternative 1 has higher utility, and takes value 0 otherwise. The additive random utility model (ARUM) specifies the utilities of alternatives 0 and 1 to be

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0 \\ U_1 &= V_1 + \varepsilon_1 \,, \end{aligned} \tag{1}$$

where V_0 and V_1 are deterministic components of utility and ε_0 and ε_1 are random components of utility. The alternative with higher utility is chosen. We observe $y = 1$, say, if $U_1 > U_0$. Due to the presence of the random components of utility this is a random event with

$$\begin{aligned} \Pr\left[y = 1 | V_0, V_1\right] &= \Pr\left[U_1 > U_0\right] \\ &= \Pr\left[V_1 + \varepsilon_1 > V_0 + \varepsilon_0\right] \\ &= \Pr\left[\varepsilon_0 - \varepsilon_1 < V_1 - V_0\right] \\ &= F\left(V_1 - V_0\right) \,, \end{aligned} \tag{2}$$

where F is the c.d.f. of $(\varepsilon_0 - \varepsilon_1)$. This yields $\Pr[y = 1] = F(\mathbf{x}'\boldsymbol{\beta})$ if $V_1 - V_0 = \mathbf{x}'\boldsymbol{\beta}$. Different choices of the functional form F generate different parametric models of binary choice (outcome).

The additive RUM model has multivariate extensions. In the general m-choice multinomial model the utility of the jth choice is specified to be given by

$$U_j = V_j + \varepsilon_j \,, \qquad j = 1, 2, \ldots, m \,, \tag{3}$$

where V_j, the deterministic component of utility may be specified to be a linear index function, e. g. $V_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ or $V_{ij} = \mathbf{x}'_i\boldsymbol{\beta}_j$, and ε_j denotes the random component of utility. Suppressing the individual subscript i for simplicity, using algebraic manipulations similar to those for the binary case, we obtain

$$\begin{aligned} \Pr[y = j] &= \Pr\left[U_j \ge U_k \,, \text{ all } k \ne j\right] \\ &= \Pr\left[\widetilde{\varepsilon}_{kj} \le -\widetilde{V}_{kj} \,, \text{ all } k \ne j\right] \,, \end{aligned} \tag{4}$$

where the tilda and second subscript j denotes differencing with respect to reference alternative j.

Consider an individual choosing a mode of transport to work where the choice set consists of train, bus, or priavte car. Each mode has associated with it a deterministic utility that depends upon attributes (e. g. money cost, time cost) of the mode and a random idiosyncratic component ("error"). Empirically the goal is to model conditional choice probabilities in terms of the mode attributes. Different multinomial models can be generated by different assumptions about the joint distribution of the error terms. These models are valid statistically, with probabilities summing to one. Additionally they are consistent with standard economic theory of rational decision-making. The idiosyncratic components of choice should exhibit correlation across choices if the alternatives are similar. For example, if the random components have independent type I extreme value distributions (a strong assumption!), then

$$\Pr[y = j] = \frac{\mathrm{e}^{V_j}}{\mathrm{e}^{V_1} + \mathrm{e}^{V_2} + \cdots \mathrm{e}^{V_m}} \,. \tag{5}$$

This is the conditional logit (CL) model when $V_j = \mathbf{x}'_j\boldsymbol{\beta}$, which means that attributes vary across choices only, and the multinomial legit (MNL) when $V_j = \mathbf{x}'\boldsymbol{\beta}_j$, which means that attributes are individual- but not choice-specific. Assuming that the random components have a joint multivariate normal distribution, which permits idiosyncratic components of utility to be correlated, generates the multinomial probit (MNP) model. The MNL is a special case of the Luce [59] model; it embodies an important structural restriction that the odds ratio for pair (i, j), $\Pr[y = i]/\Pr[y = j]$, is independent of all other available alternatives IIA). The MNP is the less restrictive Thurstone model, which allows for dependence between choices.

Multinomial Logit and Extensions The MNL model is much easier to compute than the MNP, but there is motivation for extending the MNL to allow for dependence in choices. One popular alternative is based on the generalized extreme value (GEV) model proposed by McFadden et al. [70], which leads to the nested logit (NL) model.

The GEV distribution is

$$F(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m) = \exp\left[-G\left(\mathrm{e}^{-\varepsilon_1}, \mathrm{e}^{-\varepsilon_2}, \ldots, \mathrm{e}^{-\varepsilon_m}\right)\right]$$

where the function $G(Y_1, Y_2, \ldots, Y_m)$is chosen to satisfy several assumptions that ensure the joint distribution and resulting marginal distributions are well-defined.

If the errors are GEV distributed then an explicit solution for the probabilities in the RUM can be obtained, with

$$p_j = \Pr[y = j] = \mathrm{e}^{V_j} \frac{G_j\left(\mathrm{e}^{-V_1}, \mathrm{e}^{-V_2}, \ldots, \mathrm{e}^{-V_m}\right)}{G\left(\mathrm{e}^{-V_1}, \mathrm{e}^{-V_2}, \ldots, \mathrm{e}^{-V_m}\right)}\,, \quad (6)$$

where $G_j(Y_1, Y_2, \ldots, Y_m) = \partial G(Y_1, Y_2, \ldots, Y_m)/\partial Y_j$, see McFadden (p. 81 in [70]). A wide range of models can be obtained by different choices of $G(Y_1, Y_2, \ldots, Y_m)$.

The nested logit model of McFadden [70] arises when the error terms ε_{jk} have the GEV joint cumulative distribution function

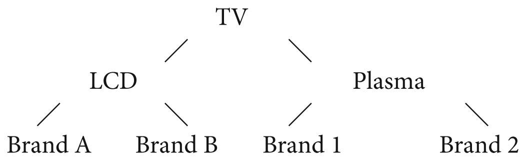
$$F(\boldsymbol{\varepsilon}) = \exp\left[-G\left(\mathrm{e}^{-\varepsilon_{11}}, \ldots, \mathrm{e}^{-\varepsilon_{1K_1}}; \ldots ; \mathrm{e}^{-\varepsilon_{J1}}, \ldots, \mathrm{e}^{-\varepsilon_{JK_J}}\right)\right] \quad (7)$$

for the following particular specification of the function $G(\cdot)$,

$$G(\mathbf{Y}) = G\left(Y_{11}, \ldots, Y_{1K_1}, \ldots, Y_{J1}, \ldots, Y_{JK_J}\right) = \sum_{j=1}^{J}\left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j}\right)^{1-\rho_j}. \quad (8)$$

The parameter ρ_j is a function of the correlation between ε_{jk} and ε_{jl} (see [13], p. 509).

The nested logit model specifies choice-making as a hierarchical process. A simple example is to consider choice of a television, where one first decides whether to buy a LCD screen or a plasma screen, and then conditional on that first choice which brand.



The random components in an RUM are permitted to be correlated for each option within the LCD and plasma groups, but are uncorrelated across these two groups. The GEV model can be estimated recursively by fitting a sequence of MNL models.

Multinomial Probit Another way to remove the IIA restriction is to assume that the unobserved components have a joint multivariate normal distribution. Beginning with m-choice multinomial model, with utility of the jth choice given by $U_j = V_j + \varepsilon_j, j = 1, 2, \ldots, m$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}]$, where the $m \times 1$ vector $\boldsymbol{\varepsilon} = [\varepsilon_1 \ldots \varepsilon_m]'$.

If the maximum likelihood equations have a unique solution for the parameters of interest, the model is said to be identified. In case that the number of equations is insufficient to yield unique estimates, restrictions on $\boldsymbol{\Sigma}$ are needed to ensure *identification*. Bunch [11] demonstrated that all but one of the parameters of the covariance matrix of the errors $\varepsilon_j - \varepsilon_1$ is identified. This can be achieved if we normalize $\varepsilon_1 = 0$, say, and then restrict one covariance element. Additional restrictions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\beta}$ may be needed for successful application, especially in models where there are no alternative-specific covariates [47]. That is, even when a MNP model is technically identified, the identification may be fragile in some circumstances, thus requiring further restrictions.

A natural estimator for this model is maximum likelihood. But, as mentioned in Sect. "Introduction", this poses a computational challenge as there is no analytical expression for the choice probabilities. For example, when $m = 3$,

$$p_1 = \Pr[y = 1] = \int_{-\infty}^{-\widetilde{V}_{31}} \int_{-\infty}^{-\widetilde{V}_{21}} f\left(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31}\right) \mathrm{d}\widetilde{\varepsilon}_{21}\, \mathrm{d}\widetilde{\varepsilon}_{31}\,,$$

where $f(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31})$ is a bivariate normal with as many as two free covariance parameters and $\widetilde{V}_{21}$ and $\widetilde{V}_{31}$ depend on regressors and parameters $\boldsymbol{\beta}$. This bivariate normal integral can be quickly evaluated numerically, but a trivariate normal integral is the limit for numerical methods. In practice it is rare to see MNP applied when there are more than 4 choices.

Simulation methods are a potential solution for higher dimensional models [89]. For Monte Carlo integration over a region of the multivariate normal, a very popular smooth GHK simulator simulator is the GHK simulator, due to Geweke [25], Hajivassiliou et al. [29] and Keane [48]; see Train [89] for details. This discussion takes $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as given but in practice these are estimated. The maximum simulated likelihood estimator (MSL) maximizes

$$\widehat{\mathrm{L}}_N(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij}\,,$$

where the $\widehat{p}_{ij}$ are obtained using the GHK or other simulator. Consistency requires the number of draws in the simulator $S \to \infty$ as well as $N \to \infty$. The method is very burdensome, especially in high dimensions. This increases the appeal of alternative estimation procedures such as the method of simulated moments (MSM). The MSM estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ solves the estimating equations

$$\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \widehat{p}_{ij})\mathbf{z}_i = \mathbf{0}\,,$$

where the $\widehat{p}_{ij}$ are obtained using an unbiased simulator. Because, consistent estimation is possible even if $S = 1$, MSM is computationally less burdensome.

Finally, Bayesian methods that exploit the latent variable structure using data augmentation approach and Markov chain Monte Carlo methods have been used successfully; see Albert and Chib [1] and McCulloch and Rossi [67].

Choice probability models are of interest on their own. More usually, however, they are of interest when linked to models of other outcomes. In observational data it is common to study outcomes that are jointly determined with the choices, often through the common dependence of the two on idiosyncratic elements. Even when the main interest is in the outcome variable, modeling of the choice component is integral to the analysis. Selection models are an example of such joint models.

Example 2: Sample Selection Models

One of the most important classes of microeconometric models is the sample selection model. Goal of modeling is usually valid inference about a target population. Sample selection problem refers to the problem of making valid inference because the sample used is not representative of the target population. Observational studies are generally based on pure random samples. A sample is broadly defined to be a selected sample if, for example, it is based in part on values taken by a dependent variable. A variety of selection models arise from the many ways in which a sample may be selected, and some of these may easily go undetected.

There is a distinction between self-selection, in which the outcome of interest is determined in part by individual choice of whether or not to participate in the activity of interest, and sample-selection, in which the participants in the activity of interest are over- or under-sampled. Selection models involve modeling the participation into the activity of interest, e. g., the labor force. The outcomes of those who participate can be compared with those of nonparticipants, which generates the counterfactual of interest. Generating and comparing counterfactuals is a fundamental aspect of selection models. Elsewhere this topic of counterfactual analysis is called treatment evaluation. When treatment evaluation is based on observational data, issues of sample selection and self-selection almost always arise.

In the example given below, consistent estimation relies on relatively strong distributional assumptions, whereas the modern trend is to do so under weaker assumptions. The example illustrates several features of microeconometric models; specifically, the model is mixed discrete-continuous and involves truncation and latent variables.

Let y_2^* denote the outcome of interest that is observed if $y_1^* > 0$. For example, y_1^* determines whether or not to work (participation) and y_2^* determines how many hours to work (outcome). The bivariate sample selection model has a participation equation,

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \le 0\,, \end{cases} \tag{9}$$

and an outcome equation,

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \le 0\,. \end{cases} \tag{10}$$

This model specifies that y_2 is observed when $y_1^* > 0$, possibly taking a negative value, while y_2 need not take on any meaningful value when $y_1^* \le 0$.

The standard specification of the model is a linear model with additive errors for the latent variables, so

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2\,, \end{aligned} \tag{11}$$

where problems arise in estimating $\boldsymbol{\beta}_2$ if ε_1 and ε_2 are correlated. If β_2 were estimated using a regression of y_2 on $\mathbf{x}_2$ using only the part of the sample for which $y_2 = y_2^*$, the resulting estimates would suffer from sample selection bias. The classic early application of this model was to labor supply, where y_1^* is the unobserved desire or propensity to work, while y_2 is actual hours worked. Heckman [33] used this model to illustrate estimation given sample selection. A popular parametric specification assumes that the correlated errors are joint normally distributed and homoskedastic, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right]. \tag{12}$$

which uses the normalization $\sigma_1^2 = 1$ because y_1^* is a latent variable that needs a measurement scale. Under general assumptions, and not just bivariate normality, the bivariate sample selection model therefore has likelihood function

$$L = \prod_{i=1}^{n} \left\{\Pr\left[y_{1i}^{*} \leq 0\right]\right\}^{1-y_{1i}} \left\{f\left(y_{2i} \middle| \, y_{1i}^{*} > 0\right) \times \Pr\left[y_{1i}{}^{*} > 0\right]\right\}^{y_{1i}} , \quad (13)$$

where the first term is the contribution when $y_{1i}^{*} \leq 0$, since then $y_{1i} = 0$, and the second term is the contribution when $y_{1i}^{*} > 0$. The model is easily estimated if it is specialized to the linear models with joint normal errors, see Amemiya [3]. An important component of the identification strategy is the use of exclusion restriction(s). This refers to the restriction that some component(s) of $\mathbf{x}_1$ affects the choice variable y_1 only, and not the outcome variable. The intuition is that this provides a source of independent variation in y_1 that can robustly identify the parameters in the y_2-equation.

The maximum likelihood approach to the estimation of self-selection models can be extended to the polychotomous choice with m-alternatives by first specifying a parametric model for choice probability that takes the form of a multinomial or nested logit, or multinomial probit, and then specifying a joint distribution between the outcome of interest and the choice probabilities; see, for example, Dubin and McFadden [20]. While straight-forward in principle, this approach does pose computational challenges. This is because analytic expressions for such joint distributions are in general not available. The problem can be addressed either by using simulation-based methods or by taking a semi-parametric formulation that permits two-step estimation of the model parameters. This topic is discussed further in Sect. "Causal Modeling".

Manski [61] and Heckman [31] were early advocates of flexible semi-parametric estimation methods, of which the "two-step Heckman procedure" is a leading example. This influential modern approach seeks to avoid strong distributional and functional form assumptions and yet obtain consistent estimates with high efficiency within this class of estimators. Following in that tradition, there is a large literature, surveyed in Lee [56], that follows the semi-parametric approach. As the dependence between choices and outcomes are central to the issue, semi-parametric IV estimators are a natural choice. One strand of the literature, represented by Blundell and Powell [9], approaches this issue form a general semiparametric IV viewpoint, whereas another, represented by Lee [58] approaches this from the perspective of linear simultaneous equations viewpoint. Whereas the latent variable approach dominates discrete choice and selection models, some econometricians, e. g. Manski [62], espouse a less restrictive model that uses the basic probability formulation of the problem, with little other structure, that can still deliver informative bounds on some counterfactual outcomes. (There are also other econometric contexts in which the bounds approach can be applied; see [63].)

Causal Modeling

An important motivation for microeconometrics stems from issues of public policies that address social and economic problems of specific groups whose members react to policies in diverse ways. Then microeconometric models are used to evaluate the impact of policy. A leading example is the effect of training on jobless workers as defined in terms of their post-training wage. Accordingly, an important topic in microeconometrics is treatment evaluation. The term treatment refers to a policy and the analogy is with the model of a clinical trial with randomized assignment to treatment. The goal is to estimate the average effect of the treatment.

Heckman [35] has pointed out that there are two types of policy evaluation questions. The first type seeks to evaluate the effect of an existing program or policy on participants relative to an alternative program or no program at all, i. e. a treatment effect. The second formulation addresses a more difficult and ambitious task of evaluating the effect of a new program or policy for which there are no historical antecedents, or of an existing program in a new economic environment. A basic tenet of econometric modeling for policy analysis is that a structural model is required to address such policy issues.

As to how exactly to define a structural model is a difficult and unsettled issue. Indeed it is easier to say what structural models are not than to define what they are. Some modelers define structural models as those that identify parameters that are invariant with respect to policies themselves; others define structural models as those that involve mathematical-statistical relationships between jointly dependent variables, and yet others define them as relationships based on dynamic optimizing models of economic behavior that embody "fundamental" taste, technology and preference parameters.

In the next section I shall provide an overview of three major approaches to causal modeling in microeconometrics. Three dominant approaches are based on, respectively, moment conditions, the potential outcome model, and the dynamic discrete choice approach.

Structural Modeling

Broadly, structural model refers to causal rather than associative modeling. Cameron and Trivedi [13] provide a definition of a structure that is based on the distinction between exogenous variables **Z**, that are taken by the modeler as given, and endogenous variables **Y**, that the modeler attempts to explain within the model; this distinction derives from the classic Cowles Commission approach for the dynamic linear SEM mentioned earlier. The dynamic linear structural SEM specifies a complete model for G endogenous variables, specified to be related to K exogenous a pre-determined variables (e. g. lagged values of **Y**).

Accordingly, a structure consists of

1. a set of variables **W** ("data") partitioned for convenience as [**Y**~**Z**];
2. a joint probability distribution of **W**, $F(\mathbf{W})$;
3. an a priori ordering of **W** according to hypothetical cause and effect relationships and specification of a priori restrictions on the hypothesized model;
4. a parametric, semiparametric or nonparametric specification of functional forms and the restrictions on the parameters of the model.

Suppose that the modeling objective is to explain the values of observable vector-valued variable **y**, $\mathbf{y}' = (y_1, \ldots, y_G)$, whose elements are functions of some other elements of **y**, and of explanatory variables **z** and a purely random disturbance u. Under the exogeneity assumption interdependence between elements of **z** is not modeled. The ith observation satisfies the set of implicit equations

$$\mathbf{g}\left(\mathbf{w}_i, \mathbf{u}_i | \boldsymbol{\theta}_0\right) = \mathbf{0}\,, \tag{14}$$

where **g** is a known function. By the Cameron–Trivedi definition this is a structural model, and to $\boldsymbol{\theta}_0$ is the vector of structural parameters. This corresponds to point 4 given earlier in this section. If there is a unique solution for $\mathbf{y}_i$ for every $(\mathbf{z}_i, \mathbf{u}_i)$, i. e.

$$\mathbf{y}_i = \mathbf{f}\left(\mathbf{z}_i, \mathbf{u}_i | \boldsymbol{\pi}\right)\ , \tag{15}$$

then this is referred to as the reduced form of the structural model, where $\boldsymbol{\pi}$ is a vector of reduced form parameters that are functions of $\boldsymbol{\theta}$. The reduced form is obtained by solving the structural model for the endogenous variables $\mathbf{y}_i$, given $(\mathbf{z}_i, \mathbf{u}_i)$. The reduced form parameters $\boldsymbol{\pi}$ are functions of $\boldsymbol{\theta}$. If the objective of modeling is inference about elements of $\boldsymbol{\theta}$, then (14) provides a direct route. Estimation of systems of equations like (14) is referred to as structural estimation in the classic Cowles Commission approach; see Heckman [34]. When the object of modeling is conditional prediction, the reduced form model is relevant.

Moment Condition Models

The classic causal model is a moment-condition model, derived from such a framework, consists of a set of r moment conditions of the form

$$E[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}\,, \tag{16}$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $\mathbf{g}(\cdot)$ is an $r \times 1$ vector function with $r \geq q$ and $\boldsymbol{\theta}_0$ denotes the value of $\boldsymbol{\theta}$ in the data generating process (d.g.p). The vector **w** includes all observables including, where relevant, a dependent (possibly vector-valued) variable **y**, potentially endogenous regressors **x** and exogenous variables **z**. The expectation is with respect to all stochastic components of **w** and hence **y**, **x** and **z**.

Estimation methods for moment condition models include fully parametric approaches such as maximum likelihood as well as semi-parametric methods such as the generalized method of moments (GMM) and instrumental variables (IV).

To make valid econometric inference on $\boldsymbol{\theta}$, it must be assumed or established that this parameter is identifiable; see Heckman [34] and Manski [62]. In other words, it is assumed that there is no set of observationally equivalent moment conditions. Identification may be established using (strong) parametric restrictions or using (weaker) semiparametric restrictions. The latter approach is currently favored in theoretical work. Point identification was emphasized in the classic Cowles Foundation but partial identification in many situations may be more attainable, especially if weaker restrictions on probability distributions of data are used; see Manski [63]. However, assuming point identification and given sufficient data, in principle these moment conditions lead to a unique estimate of the parameter $\boldsymbol{\theta}$. Potentially there are many reasons for loss of identifiability. Some of these are discussed in the next section where we also consider identification strategies.

The above approach has limitations. First, the definition of structure is not absolute because the distinction between endogenous **Y** and exogenous **Z** may be arbitrary. Second, the parameters $\boldsymbol{\theta}$ need not be tied to fundamental (or "deep") parameters; indeed it includes both the policy parameters that are of intrinsic interest and others that are not. If, however, the moment conditions are derived either from a model of optimization, or from some fundamental postulates of economic behavior such as the efficient market hypothesis, then at least some subset of parameters $\boldsymbol{\theta}$ can have a "structural" interpretation that is based on preference or technology parameters. Some econometricians prefer a narrower definition of a causal parameter which focuses only on the impact of the policy on the outcome of interest; the remaining parameters are treated as non-causal. Third, the approach is often difficult to implement in a way that provides information about either of the types of policy issues mentioned at the beginning of this section.

In response to these difficulties of the conventional approach two alternative approaches have emerged. The first is the potential outcome model (POM) that can be historically traced back to Neyman and Fisher. The second (and more modern) approach is based on dynamic stochastic Markov models. The first is easier to implement and hence currently dominates the applied literature. Next I will provide a brief overview of each approach.

Treatment Effect Models

This section deals with two closely related approaches in the treatment evaluation literature which targets an important structural parameter and its variants. Treatment effect models have been used extensively to study, to give just a few examples, the effect of: schooling on earnings, the class size on scholastic performance, unions on wages, and health insurance on health care use. Although in many cases the treatment variable is dichotomous, the framework can handle polychotomous treatment variables also. Treatment need not be discrete; the framework can handle ordered as well as continuously varying treatments.

Potential Outcome Models Much econometric estimation and inference are based on observational data. Identification of and inference on causal parameters is very challenging in such a modeling environment. Great simplification in estimating causal parameters arise if one can use data from properly designed and implemented controlled social experiments. Although such experiments have been implemented in the past they are generally expensive to organize and run. Econometricians therefore seek out data generated by quasi- or natural experiments which may be thought of as settings in which some causal variable changes exogenously and independently of other explanatory variables. This is an approximation to a controlled trial.

Random assignment implies that individuals exposed to treatment are chosen randomly, and hence the treatment assignment does not depend upon the outcome and is uncorrelated with the attributes of treated subjects. The great resulting simplification in relating outcomes to policy changes is unfortunately rarely achievable because random assignment of treatment is generally not feasible in economics. Most analyzes have to depend upon observational data.

As an example, suppose one wants to study the effect of unions on wages using data from unionized and nonunionized workers. Here being a unionized worker is the treatment. For the unionized worker, being a nonunion worker is the counterfactual. The purpose of the causal model is to estimate the mean difference in wages of unionized and nonunionized workers, the difference being attributed to being in the union.

A major obstacle for causality modeling stems from the so-called fundamental problem of causal inference [40]. Accordingly, in an observational setting one can only observe an individual in either the treated or the untreated state, and not both. Hence one cannot directly observe the effect of the treatment. Consequently, nothing more can be said about causal impact without some hypothesis about the counterfactual, i. e. what value of the outcome would have been observed in the absence of the change in policy variable.

The POM, also known as the Rubin causal model (RCM), provides a solution to the problem of establishing a counterfactual for policy evaluation. Causal parameters based on counterfactuals provide statistically meaningful and operational definitions of causality. In the POM framework the term "treatment" is used interchangeably with "cause". All policy changes and changes in the policy environment are broadly covered by the term treatment. Given a group impacted by policy, and another one that is not, a measure of causal impact is the average difference in the outcomes of the treated and the nontreated groups. Examples of treatment-outcome pairs are: health insurance and health care utilization; schooling and wages; class size and scholastic performance. Of course, the fact that with observational data a treatment is often chosen, not randomly assigned, is a significant complication.

In the POM framework, assuming that every element of the target population is potentially exposed to the treatment, the variables $(y_{1i}, y_{0i}, D_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, forms the basis of treatment evaluation. The categorical variable D takes the values 1 and 0, respectively when treatment is or is not received; y_{1i} measures the response for individual i receiving treatment, and y_{0i} when not receiving treatment, $\mathbf{x}_i$ is the vector of exogenous covariates. That is, $y_i = y_{1i}$ if $D_i = 1$ and $y_i = y_{0i}$ if $D_i = 0$. Receiving and not receiving treatment are mutually exclusive states so only one of the two measures is available for any given i; the unavailable measure is the counterfactual. The effect of the cause D on outcome if individual i is measured by $(y_{1i} - y_{0i})$. The average causal effect of $D_i = 1$, relative to $D_i = 0$, is measured by the average treatment effect

(ATE):

$$\text{ATE} = E[y|D = 1, \mathbf{x}] - E[y|D = 0, \mathbf{x}] \,, \tag{17}$$

where expectations are with respect to the probability distribution over the target population. Unlike the conventional structural model that emphasizes marginal effects the POM framework emphasizes ATE and parameters related to it.

POM can lead to causal statements if the counterfactual can be clearly stated and made operational. In observational data, however, a clear distinction between observed and counterfactual quantities may not be possible. Then ATE will estimate a weighted function of the marginal responses of specific subpopulations. Despite these difficulties, the identifiability of the ATE parameter may be an easier research target.

Matching Methods In the POM framework a causal parameter may be unidentified because there is no suitable comparison or control group that provides the benchmark for estimation. In observational studies, by definition there are no experimental controls. Therefore, there is no direct counterpart of the ATE calculated as a mean difference between the outcomes of the treated and nontreated groups. In other words, the counterfactual is not identified.

Matching methods provide a potential solution by creating a synthetic sample which includes a comparison group that mimics the control group. Such a sample is created by matching. Potential comparison units, that are not necessarily drawn from the same population as the treated units, are those for whom the observable characteristics, $\mathbf{x}$, match those of the treated units up to some selected degree of closeness. In the context of the unionization example, one would match, as closely as possible, unionized with nonunionized workers in terms of a vector of observable characteristics. Of course, if there are significant unobserved sources of differences that cannot be controlled, then this could lead to omitted variable bias. Given a treated sample plus well matched controls, under certain assumptions it becomes possible to identify parameters related to the ATE.

Matching may produce good estimates of the average effect of the treatment on the treated, i. e. the ATET parameter if (1) we can control for a rich set of $\mathbf{x}$ variables, (2) there are many potential controls. It also requires that treatment does not indirectly affect untreated observations. The initial step of establishing the nearest matches for each observation will also clarify whether comparable control observations are available.

Suppose the treated cases are matched in terms of all observable covariates. In a restricted sense all differences between the treated and untreated groups are controlled. Given the outcomes y_{1i} and y_{0i}, for the treatment and control, respectively, the average treatment effect is

$$\begin{aligned} &E\left[y_{1i}|D_i = 1\right] - E\left[y_{0i}|D_i = 0\right] \\ &\quad = E[y_{1i} - y_{0i}|D_i = 1] + \left\{E\left[y_{0i}|D_i = 1\right] \right. \\ &\qquad\qquad \left. - E\left[y_{0i}|D_i = 0\right]\right\} \,. \end{aligned} \tag{18}$$

The first term in the second line is the ATET, and the second bracketed term is a "bias" term which will be zero if the assignment to the treatment and control is random. The sample estimate of ATET is a simple average of the differential due to treatment.

There is an extensive literature on matching estimators covering both parametric and nonparametric matching estimators; see Lee [58] for a survey. Like the POM framework, the approach is valid for evaluating policy that is already in operation and one that does not have general equilibrium effects. An important limitation is that the approach is vague and uninformative about the mechanism through which the treatment effects occur.

Dynamic Programming Models

Dynamic programming (DP) models represent a relatively new approach to microeconometric modeling. It emphasizes structural estimation and is often contrasted with "atheoretical" models that are loosely connected to the underlying economic theory. The distinctive characteristics of this approach include: a close integration with underlying theory; adherence to the assumption of rational optimizing agents; generous use of assumptions and restrictions necessary to support that close integration; a high level of parametrization of the model; concentration on causal parameters that play a key role in policy simulation and evaluation; and an approach to estimation of model parameters that is substantially different from the standard approaches used in estimating moment condition models. The special appeal of the approach comes from the potential of this class of models to address issues relating to new policies or old policies in a new environment. Further, the models are dynamic in the sense that they can incorporate expectational factors and inter-temporal dependence between decisions.

There are many studies that follow the dynamic programming approach. Representative examples are Rust [81]; Hotz and Miller [42]; Keane and Wolpin [50]. Some key features of DP models can be exposited using a model due to Rust and Phelan [85] which provides an empirical analysis of how the incentives and constraints of the US social security and Medicare insurance system affects the labor supply of old workers. Some of the key constraints arise due to incomplete markets, while individual behavior is based in part on expectations about future income streams. Explaining transitions from work to retirement is a challenging task not only because it involves forward-looking behavior in a complex institutional environment but also because a model of retirement behavior must also capture considerable heterogeneity in individual labor supply, discontinuities in transitions from full time work to not working, and presence of part-time workers in the population, and coordination between labor supply decisions and retirement benefits decisions.

The main components of the DP model are as follows. State variable is denoted by s_t, control variable by d_t. β is the intertemporal discount factor. In implementation all continuous state variables are discretized – a step which greatly expands the dimension of the problem. Hence all continuous choices become discrete choices, d_t is a discrete choice sequence, and the choice set is finite. For example, in Rust and Phelan [85] total family income is discretized into 25 intervals, social security state into 3 states, and employment state (hours worked annually) into 3 discrete intervals, and so forth. There is a single period utility function $u_t(s, d, \theta_u)$ and $p_t(s_{t+1}|s_t, d_t, \theta_p, \alpha)$ denotes the probability density of transitions from s_t to s_{t+1}. The optimal decision sequence is denoted by $\delta = (\delta_0, \ldots, \delta_T)$ where $d_t = \delta_t(s_t)$ and is the optimal solution that maximizes the expected discounted utility:

$$V_t(s) = \max_{\delta} E_{\delta} \left\{ \sum_{j=t}^{T} \beta^{j-t} u_j(s_j, d_j, \theta_u) | s_t = s \right\} . \quad (19)$$

The model takes Social Security and Medicare policy parameters, α, as known. The structural parameters $\theta = (\beta, \theta_u, \theta_p)$ are to be estimated. To specify the stochastic structure of the model the state variables are partitioned as $s = (x, \eta)$, where x is observable and η is unobservable (for the econometrician); $\eta_t(d)$ can be thought of as the net utility or disutility impact due to factors unobserved by the econometrician at time t.

An important assumption, due to Rust [81], which restricts the role of η permits the following decomposition of the joint probability distribution of (x_{t+1}, η_{t+1}):

$$\Pr\left[x_{t+1}, \eta_{t+1}|x_t, \eta_t, d_t\right] = \Pr\left[\eta_{t+1}|x_{t+1}\right] \Pr\left[x_{t+1}|x_t, d_t\right] .$$

Note that the first term on the right-hand side implies serial independence of unobservables; the second term has a Markov structure and implies that η_t affects x_t only through d_t.

$$v_t(x_t, d_t, \theta, \alpha) = u_t(x_t, d_t, \theta_u) + \beta \int \log \left[\sum_{d_{t+1} \in D(x_{t+1})} \exp\{v_{t+1}(x_{t+1}, d_{t+1}, \theta, \alpha)\} \right] p_t(x_{t+1}|x_t, d_t, \theta_p, \alpha) , \quad (20)$$

Estimation of the model, based on panel data $\{x_t^i, d_t^i\}$, uses the likelihood function

$$L(\theta) = L(\beta, \theta_u, \theta_p) = \prod_{i=1}^{I} \prod_{t=1}^{T_i} P_t(d_t^i|x_t^i, \theta_u) p_t(x_t^i|x_{t-1}^i, d_{t-1}^i, \theta_p) . \quad (21)$$

This is a high dimensional model because a large number of state variables and associated parameters are needed to specify the future expectations. (This complexity is highlighted to emphasize that DP models run into dimensionality problems very fast.) First, strong assumptions are needed to address the unobservable and subjective aspects of decision-making because there are a huge number of possible future contingencies to take into account. Second, restrictions are needed to estimate the belief arrays. Consistent with tenets of rational agents the model assumes rational expectations. To impose exclusion restrictions p_t is decomposed into a product of marginal and conditional densities.

As a simplification a two-stage estimation procedure is used: (1) estimate θ_p using first stage partial likelihood function involving only products of the p_t terms; (2) estimate θ_p by solving the DP problem numerically, and estimate (β, θ_u) using a second stage partial likelihood function consisting of only products of P_t. The two-stage estimation procedure is not as efficient as the full maximum likelihood estimation since the error in $\hat{\theta}_p$ contaminates the estimated covariance matrix for θ_u.

Space limitations do not permit us to provide the details of the computational procedure, for which we refer the reader to Rust [81]. In outline, at the first step the procedure estimates the transition probability parameters θ_p using the partial likelihood function and at the second stage a Nested Fixed Point (NFP) algorithm is used to estimate the remaining parameters.

New Directions in Structural Modeling

The motivation for many of the recent developments lies in the difficulties and challenges of identifying causal parameters under fewer distributional and functional form restrictions. Indeed an easily discernible trend in modern research is steady movement away from strong parametric models and towards semi-parametric models. Increasingly semiparametric identification is the stated goal of theoretic research [41]. Semiparametric identification means that unique estimates of the relevant parameters can be obtained without making assumptions about distribution of data, and some times it also means that assumptions about functional forms can also be avoided. Potentially there are numerous ways in which the identification of key model parameters can be compromised. The solution strategy in such cases is often model specific. This section provides a selective overview of recent developments in microeconometrics that address such issues.

Endogeneity and Multivariate Modeling

Structural nonlinear models involving LDVs arise commonly in microeconometrics. A leading example of a causal model involves modeling the conditional distribution (or moments) of a continuous outcome (y) which depends on variables ($\mathbf{x}$, D) where D is an endogenous binary treatment variable. For example, y is medical expenditure and D is a binary indicator of health insurance status. The causal parameter of interest is the marginal effect of D on y. More generally y could be binary, count, an ordered discrete variable, or a truncated/censored continuous variable. More generally the issue is that multivariate modeling. Currently there is no consensus on econometric methodology for handling this class of problems. Some of the currently available approaches are now summarized.

Control Functions A fully parametric ("full information") estimation strategy requires the specification of the joint distribution of (y, D), which is often difficult because such a joint distribution is rarely available. Another ("limited information") strategy is to estimate only the conditional model, quite often only the conditional mean $\mathrm{E}[y|\mathbf{x}, D]$, controlling for endogeneity of treatment. If the model is additively separable in $\mathrm{E}[y|\mathbf{x}, D]$ and the stochastic error ε which is correlated with d so that $\mathrm{E}(\varepsilon D) \neq 0$, then a two-step procedure may be used. This involves replacing D by its projection on a set of exogenous instrumental variables $\mathbf{z}$ (usually including $\mathbf{x}$), denoted $\widehat{D}(\mathbf{z})$, and estimating the conditional expectation $\mathrm{E}[y|\mathbf{x}, \widehat{D}(\mathbf{z})]$. Unfortunately, this approach does not always yield a consistent estimate of the causal parameter; for example, if the conditional mean is nonlinear in ($\mathbf{x}$, D). Therefore this approach is somewhat ad hoc.

Another similar strategy, called the control function approach, involves replacing $\mathrm{E}[y|\mathbf{x}, D]$ by $\mathrm{E}[y|\mathbf{x}, \mathbf{w}, D]$. Here $\mathbf{w}$ is a set of additional variables in the conditional mean function such that the assumption $\mathrm{E}(\varepsilon D|\mathbf{w}) = 0$; that is D can be treated as exogenous, given the presence of $\mathbf{w}$ in the conditional mean function. Again such an approach does not in general identify the causal parameter of interest. Additional restrictions are often necessary for structural identification. In a number of cases where the approach has been shown to work some functional form and structural restrictions are invoked, such as additive separability and a triangular error structure.

Consider the following example of an additively separable model with a triangular structure. Let y_1 be the dependent variable in the outcome equation, which is written as

$$y_1 = E\left[y_1|D, \mathbf{x}\right] + u_1 + \lambda u_2 \, ,$$

where $(u_1 + \lambda u_2)$ is the composite error. Let D denote the treatment indicator for which the model is

$$D = E\left[D|\mathbf{z}\right] + u_2 \, .$$

A simple assumption on the distribution of the error terms takes them to be zero-mean and mutually uncorrelated. In this case the control function approach can be used. Specifically a consistent estimate of u_2, say $\widehat{u}_2$, can be included as an additional regressor to the y_1 equation. This type of argument has been used for handling endogeneity in regression models that are specified for, instead of the conditional mean, the conditional median or conditional quantile regression; see Chesher [14], Ma and Koenker [60]. The control function approach has been adapted for treating endogeneity problems in semiparametric and nonparametric framework [9].

Latent Factor Models Another "full information" approach that simultaneously handles discrete variation and endogeneity also imposes a restriction on the structure of dependence using latent factors and resorts to simulation-assisted estimation. An example is Deb and Trivedi [18] who develop a joint model of counts, with a binary insurance plan variable (D) as a regressor, and a model for the choice of insurance plan. Endogeneity in their model arises from the presence of correlated unobserved heterogeneity in the outcome (count) equation and the binary choice equation. Their model has the following structure:

$$\begin{aligned} \Pr\left[Y_i = y_i|\mathbf{x}_i, D_i, l_{ji}\right] &= f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \, . \\ \Pr\left[D_i = 1|\mathbf{z}_i, l_{ji}\right] &= g\left(\mathbf{z}_i'\boldsymbol{\alpha} + \delta l_i\right) \, . \end{aligned}$$

Here l_i is latent factor reflecting unobserved heterogeneity and δ is an associated factor loading. The joint distribution of selection and outcome variables, conditional on the common latent factor, can be written as

$$\begin{aligned}\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i, l_i\right]\\ = f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right) ,\end{aligned} \quad (22)$$

because (y, D) are conditionally independent.

The problem in estimation arises because the l_i are unknown. Although the l_i are unknown, assume that the distribution of l_i, h, is known and can therefore be integrated out of the joint density, i. e.,

$$\begin{aligned}\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i\right]\\ = \int \left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right)\right]\\ h\left(l_i\right) \mathrm{d}l_i .\end{aligned}$$

Cast in this form, the unknown parameters of the model may be estimated by maximum likelihood.

The maximum likelihood estimator maximizes the joint likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_i, D_i, \mathbf{x}_i, \mathbf{z}_i)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \lambda)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}, \delta)$, refer to parameters in the outcome and plan choice equations respectively, and L refers to the joint likelihood.

The main problem of estimation, given suitable specifications for f, g and h, is the fact that the integral does not have, in general, a closed form solution. The maximum simulated likelihood (MSL) estimator involves replacing the expectation by a simulated sample average, i. e.,

$$\begin{aligned}\widetilde{\Pr}[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i]\\ = \frac{1}{S}\sum_{s=1}^{S}\left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \sum\nolimits_j \lambda\widetilde{l}_{is}\right)\right.\\ \left.\times g\left(z_i'\boldsymbol{\alpha} + \delta\widetilde{l}_{is}\right)\right] ,\end{aligned} \quad (23)$$

where $\widetilde{l}_{is}$ is the sth draw (from a total of S draws) of a pseudo-random number from the density h and $\widetilde{\Pr}$ denotes the simulated probability.

The above approach, developed for an endogenous dummy regressor in a count regression model, can be extended to multiple dummies (e. g. several types of health insurance), and multiple outcomes, discrete or continuous (e. g. several measures of health care utilization such as number of doctor visits, prescribed medications). The limitation comes from the heavy burden of estimation compared with an IV type estimator. Further, as in any simultaneous equation model, identifiability is an issue. Applied work typically includes some nontrivial explanatory variables in the **z** vector that are excluded from the **x** vector. As an example, consider insurance premium which would be a good predictor of insurance status but will not directly affect health care use.

Instrumental Variables and Natural Experiments

If identification is jeopardized because the treatment variable is endogenous, then a standard solution is to use valid instrumental variables. To identify the treatment effect parameter we need exogenous sources of variation in the treatment. Usually this means that the model must include at least the minimum number of exogenous variables (instruments) that affect the outcome only through the treatment – an assumption usually called an exclusion or identification restriction. This requirement may be difficult to satisfy. Keane [49] gives an example where there are no possible instruments. Even this extreme possibility is discounted, agreement on valid instruments is often difficult, and when such agreement can be established the instruments may be "weak" in the sense that they do not account for substantial variation in the endogenous variables they are assumed to affect directly. The choice of the instrumental variable as well as the interpretation of the results obtained must be done carefully because the results may be sensitive to the choice of instruments. In practice, such instrumental variables are either hard to find, or they may generate only a limited degree of variation in the treatment by impacting only a part of the relevant population.

A natural experiment may provide a valid instrument. The idea here is simply that a policy variable may exogenously change for some subpopulation while remaining unchanged for other subpopulations. For example, minimum wage laws in one state may change while they remain unchanged in a neighboring state. Such events create natural treatment and control groups. Data on twins often provide data with both natural treatment and control, as has been argued in many studies that estimate the returns to schooling; see Angrist and Krueger [5]. If the natural experiment approximates randomized treatment assignment, then exploiting such data to estimate structural parameters can be simpler than estimation of a larger simultaneous equations model with endogenous treatment variables. However, relying on data from natural experiments is often not advisable because of such events are rare and because the results from them may not generalize to a broader population.

Limitations of the IV Approach Some limitations of the IV approach, e. g. the weak IV problem, are general but certain others are of special relevance to microeconometrics. One of these is a consequence of heterogeneity in the impact of the policy on the outcome. Consideration of this complication has led to significant refinements in the interpretation of results obtained using the IV method.

In many applications of the POM framework, the underlying assumption is that there exists a comparison group and a treatment that is homogeneous in its response to the treatment. In the heterogeneous case, the change in the participation in treatment generated by the variation in the instrument may depend both upon which instrument varies, and on the economic mechanism that links the participation to the instrument. As emphasized by Heckman and Vytlacil [38], Keane [49] and others, a mechanical application of the IV approach has a certain black box character because it fails to articulate the details of the mechanism of impact. Use of different instruments identify different policy impact parameters because they may impact differently on different members of the population. Heckman and Vytlacil [38] emphasize that the presence of unobserved heterogeneity and selection into treatment may be based on unobserved gains, a condition they call *essential heterogeneity*. The implication for the choice of IVs is that these may be independent of the idiosyncratic gains in the overall population, but conditional on those who self-select into treatment, they may no longer be independent of the idiosyncratic gains in this subgroup. Further, as a consequence of the dependence between treatment choice and IV estimates different IVs identify different parameters. In this context, an a priori specification of the choice model for treatment is necessary for the interpretation of IV estimators.

The concept of local instrumental variables is related to the local average treatment (LATE) parameter introduced by Imbens and Angrist [45]. To illustrate this we consider the following canonical linear model.

The outcome equation is a linear function of observable variables **x** and a participation indicator D:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i \;, \tag{24}$$

and the participation decision depends upon a single variable z, referred to as an instrument,

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i \;, \tag{25}$$

where D_i^* is a latent variable with its observable counterpart generated by

$$D_i = \begin{cases} 0 \text{ if } D_i^* \le 0 \\ 1 \text{ if } D_i^* > 0 \end{cases} . \tag{26}$$

There are two assumptions: (1) There is an exclusion restriction as the variable z that appears in the equation for D that does not appear in the equation for **x**. (2) Conditional on $(\mathbf{x}, z)$ Cov $[z, v]$ = Cov $[u, z]$ = Cov $[\mathbf{x}, u]$ = 0, but Cov $[D, z] \neq 0$. It is straightforward to show that the IV estimator of the treatment effect parameter α is

$$\alpha_{IV} = \frac{E\left[y|z'\right] - E\left[y|z\right]}{\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right]} \;, \tag{27}$$

which is well-defined if $\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right] \neq 0$. The sample analog of α_{IV} is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated due to the change in z.

Why does this measure a "local" effect? This is because the treatment effect applies to the "compliers" only, that is those who are induced to participate in the treatment as a result of the change in z; see Angrist et al. [6]. Thus LATE depends upon the particular values of z used to evaluate the treatment and on the particular instrument chosen. Those who are impacted may not be representative of those treated, let alone the whole population. Consequently the LATE parameter may not be informative about the consequences of large policy changes brought about by changes in instruments different from those historically observed.

If more than one instrument appears in the participation equation, as when there exist overidentifying restrictions, the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

Omitted Variables, Fixed and Random Effects

Identification may be threatened by the presence of a large number of nuisance or incidental parameter. For example, in a cross section or panel data regression model the conditional mean function may involve an individual specific effect α_i, i. e. $\mathrm{E}[y_i|\mathbf{x}_i, \alpha_i]$ or $\mathrm{E}[y_{it}|\mathbf{x}_{it}, \alpha_i]$ where $i = 1, \ldots, N$, $t = 1, \ldots, T$. The parameters α_i may be interpreted as omitted unobserved factors. Two standard statistical models for handling them are fixed and random effect formulations. In a fixed-effect (FE) model the α_i are assumed to be correlated with the observed covariates $\mathbf{x}_i$, i. e. $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] \neq \mathbf{0}$, whereas in the random effects (RE) model $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] = \mathbf{0}$ is assumed. Because the FE model is less restrictive, it has considerable appeal in microeconometrics.

FE Models In FE models this effect cannot be identified without multiple observations on each individual, i. e.. panel data. Identification is tenuous even with panel data if the panel is short, i. e. T is small; see Lancaster [54] about the incidental parameters problem. The presence of these incidental parameters in the model also hinders the identification of other parameters of direct interest. A feasible solution in the case where both N and T are large, is to introduce dummy variables for each individual and estimate all the parameters. The resulting computational problem has a large dimension but has been found to be feasible not only in the standard case of linear regression but also for some leading nonlinear regressions such as the Probit, Tobit and Poisson regressions [26,27].

If the panel is short, the $\alpha_i (i = 1, \ldots, N)$ cannot be identified and no consistent estimator is available. Then the identification strategy focuses on the remaining parameters that are estimated after eliminating α_i by a transformation of the model. Consider, as an example, the linear model with both time-varying and time-invariant exogenous regressors $(\mathbf{x}'_{it}, \mathbf{z}'_i)$

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \varepsilon_{it} \,, \qquad (28)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are common parameters, while $\alpha_1, \ldots, \alpha_N$ are incidental parameters if the panel is short as then each α_i depends on fixed T observations and there are infinitely many α_i since $N \to \infty$. Averaging over T observations yields

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}'_i\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \bar{\varepsilon}_i \,. \qquad (29)$$

On subtracting we get the "within model"

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) \,, \qquad i = 1, \ldots, N \,, \quad t = 1, \ldots, T \,, \qquad (30)$$

where the α_i term and the time-invariant variables $\mathbf{z}_i$ disappear. A first difference transformation $y_{it} - y_{i,t-1}$ can also eliminate the α_i. The remaining parameters can be consistently estimated, though the disappearance of variables from the model means that prediction is no longer feasible.

Unfortunately this elimination "trick" does not generalize straight-forwardly to other models, especially nonlinear nonnormal models with fully specified distribution. There is no unified solution to the incidental parameters problem, only model-specific approaches. In some special cases the conditional likelihood approach does solve the incidental parameter problem, e. g. linear models under normality, logit models (though not probit models) for binary data, and some parametrizations of the Poisson and negative binomial models for count data. The RE model, by contrast, can be applied in more widely.

RE Panel Models If the unobservable individual effects $\alpha_i, \alpha_i > 0$, are random variables that are distributed independently of the regressors, the model is called the random effects (RE) model. Usually the additional assumptions that both the random effects and the error term are also employed, i. e., $\alpha_i \sim [\alpha, \sigma^2_\alpha]$, and $\varepsilon_{it} \sim [0, \sigma^2_\varepsilon]$ are also employed. More accurately this is simply the *random intercept model*. As a specific example consider the Poisson individual-specific effects model which specifies

$$y_{it} \sim \text{Poisson}[\alpha_i \exp(\mathbf{x}'_{it}\boldsymbol{\beta})] \,.$$

If we assume gamma distributed random effects distributed with mean 1, variance $1/\gamma = \eta$ and density $g(\alpha_i|\eta) = \eta^\eta \alpha_i^{\eta-1} \mathrm{e}^{-\alpha_i\eta}/\Gamma(\eta)$, there is a tractable analytical solution for the unconditional joint density for the ith observation $\int \left[\prod_{t=1}^{T} f(y_{it}|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\right] g(\alpha_i|\eta)\,\mathrm{d}\alpha_i$ (see Cameron and Trivedi ([13]: chapter 23.7 for algebraic details). However, under other assumptions about the distribution (e. g. log-normal) a closed form unconditional density usually does not arise, and estimation is then based on numerical integration – an outcome that is fairly common for nonlinear random effect models.

Modeling Heterogeneity

To accommodate the diversity and complexity of responses to economic factors, it is often necessary to allow for variation in the model parameters. There are many specification strategies to accomplish such a goal. One of the most popular and well-established strategy is to model heterogeneity using some type of mixture model. Typically the specification of a mixture model involves two steps. In the first step a conditional distribution function $F(y|\mathbf{x}, \boldsymbol{\nu})$ is specified where $\mathbf{x}$ is an observed vector of covariates and ν is an unobserved heterogeneity term, referred to as frailty in biostatistics. In the second step a distribution $G(\nu)$ is specified for ν and a mixture model is derived. The distribution of ν may be continuous or discrete. Poisson-gamma mixture for count data and Weibull–gamma mixtures for survival data are two leading examples based on continuous heterogeneity assumption. The mixed multinomial logit model (MMNL) is another example [75].

The mixture class of models is very broad and includes two popular subclasses, the random coefficient approach and the latent class approach. While relatively simple in formulation, such mixture approaches often generate major identification and computational challenges [24]. Here I provide two examples that illustrate the issues associated with their use.

Latent Class Models Consider the following two-component *finite mixture* model. If the sample is a probabilistic mixture from two subpopulations with p.d.f. $f_1(y|\mu_1(\mathbf{x}))$ and $f_2(y|\mu_2(\mathbf{x}))$, then $\pi f_1(.) + (1 - \pi) f_2(.)$, where $0 \leq \pi \leq 1$, defines a two-component finite mixture. That is, observations are draws from f_1 and f_2, with probabilities π and $1-\pi$ respectively. The parameters to be estimated are (π, μ_1, μ_2). The parameter π may be further parameterized.

At the simplest level we think of each subpopulation as a "type", but in many situations a more informative interpretation may be possible. There may be an a priori case for such an interpretation if there is some characteristic that partitions the sampled population in this way. An alternative interpretation is simply that the linear combination of densities is a good approximation to the observed distribution of y. Generalization to additive mixtures with three or more components is in principle straight-forward but subject to potential problems of the identifiability of the components.

Formally the marginal (mixture) distribution is

$$h\left(t_i|\mathbf{x}_i, \pi_j, \boldsymbol{\beta}\right) = \sum_{j=1}^{m} f\left(t_i|\mathbf{x}_i, \nu_j, \boldsymbol{\beta}\right) \pi_j\left(\nu_j\right) \,, \tag{31}$$

where ν_j is an estimated support point and π_j is the associated probability. This representation of unobserved heterogeneity is thought of as semiparametric because it uses a discrete mass point distribution. The specification has been found to be very versatile. It has been used to model duration data where the variable of interest is the length of time spent in some state, e. g. unemployment, and individuals are thought to differ both interms of their observable and unobservable characteristics; see Heckman and Singer [37].

The estimation of the finite mixture model may be carried out either under the assumption of known or unknown number of components. More usually the proportions $\pi_j, j = 1, \ldots, m$ are unknown and the estimation involves both the π_j and the component parameters. The maximum likelihood estimator for the latter case is called Nonparametric Maximum Likelihood Estimator (NPMLE), where the nonparametric component is the number of classes. Estimation is challenging, especially if m is large because the likelihood function is generally multi-modal and gradient-based methods have to be used with care. If the number of components is unknown, as is usually the case, then some delicate issues of inference arise. In practice, one may consider model comparison criteria to select the "best" value of m. Baker and Melino [8] provide valuable practical advice for choosing this parameter using an information criterion.

LC models are very useful for generating flexible functional forms and for approximating the properties of nonparametric models. For this reason it has been used widely. Deb and Trivedi [16,17] use the approach for modeling mixtures of Poisson and negative binomial regressions. McFadden and Train [75] show that latent class multinomial logit model provides an arbitrarily good approximation to any multinomial choice model based on the RUM. This means that it provides one way of handling the IIA problem confronting the users of the MNL model. Dynamic discrete choice models also use the approach.

LC models generate a computational challenge arising from having to choose m and to estimate corresponding model parameters for a given m, and there is the model selection problem. Often there is no prior theory to guide this choice which in the end may be made largely on grounds of model goodness-of-fit. Akaike's or Bayes penalized likelihood (or information) criterion (AIC or BIC) is used in preference to the likelihood ratio test which is not appropriate because of the parameter boundary hypothesis problem. The dimension of parameters to be estimated is linear in m, the number of parameters can be quite large in many microeconometric applications that usually control for many socio-demographic factors. This number can be decreased somewhat if some elements are restricted to be equal, for example by allowing the intercept but not the slope parameters to vary across the latent classes; see, for example, Heckman and Singer [37].

When the model is overparametrized, perhaps because the intergroup differences are small, the parameters cannot be identified. The problem is reflected in slow convergence in computation due to the presence of multiple optima, or a flat likelihood surface. The computational algorithm may converge to different points depending on the starting values.

Interpretation of the LC model can be insightful because it has the potential to capture diverse responses to different stimuli. However, a potential limitation is due to the possibility that additional components may simply reflect the presence of outliers. Though this is not necessarily a bad thing, it is useful to be able to identify the outlying observations which are responsible for one or more components.

Random Coefficient Models Random coefficient (RC) models provide another approach to modeling heterogeneity. The approach has gained increasing popularity especially in the applications of discrete choice modeling to marketing data. In this section I provide an exposition of the random coefficient logit model based on Train [89] where a more comprehensive treatment is available. The random coefficient models extend the RUM model of Sect. "Introduction" which restricts the coefficients of parameters to be constant across individuals. If individuals have different utility functions then that is a misspecification. The RC framework is one of a number of possible ways of relaxing that restriction.

The starting point is the RUM framework presented in Sect. "Introduction". Assume individual $i(i = 1, 2, \ldots, N)$ maximizes utility U_{ij} by choosing alternative j from her choice set $M_n = (0, 1)$. The utility U_{nj} has observed (systematic) part $V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i)$ and random part ε_{ij};

$$U_{ij} = V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i) + \varepsilon_{ij} \,, \qquad j = 0, 1 \,; \quad i = 1, 2, \ldots, N \,. \tag{32}$$

Vector $\mathbf{X}_{nj}$ in $V(.,.)$ represents observed attributes of alternatives, characteristics of the individual i as well as alternative-specific constants. $\boldsymbol{\beta}_i$ is the vector of coefficients associated with $\mathbf{X}_{ij}$. Error term ε_{ij} captures unobserved individual characteristics/unobserved attributes of the alternative j and follows some distribution $\mathrm{D}(\boldsymbol{\theta}_\varepsilon)$, where $\boldsymbol{\theta}_\varepsilon$ is the unknown parameter vector to be estimated. Of course, U_{ij} is latent, so we use an indicator function, y_{ij}, such that $y_{ij} = 1$ if $U_{ij} \geq U_{ik} \forall k \neq j$ and $y_{ij} = 0$ otherwise. Probability that individual i chooses alternative j is

$$\begin{aligned} P_{ij} = P(j|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P(y_{ij} = 1) \\ = P(U_{ij} \geq U_{ii} \forall i \neq j) \,, \end{aligned}$$

and the probability that alternative j is chosen is $P(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P_{ij}^{y_{ij}}$, which, under the independence assumption, leads to the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}_\varepsilon) = \prod_{i=1}^{N} \prod_{j \in M_i} P_{ij}^{y_{ij}} \,. \tag{33}$$

Different assumptions on the error structure lead to different discrete choice models. The kth component of the vector $\boldsymbol{\beta}_i$, which represents the coefficient of some attribute k, can be decomposed as $\beta_{ik} = b + \boldsymbol{\delta}' \boldsymbol{\omega}_i + \sigma_k \eta_{ik}$, if the coefficient is random and simply $\beta_{nk} = b$, if the coefficient is non-random. Here b represents the average taste in the population for provider attribute k, $\boldsymbol{\omega}_i$ is a vector of choice-invariant characteristics that generates individual heterogeneity in the means of random coefficients $\boldsymbol{\beta}_i$, and $\boldsymbol{\delta}$ is the relevant parameter vector. Finally, η_{ik} is the source of random taste variation, which is be assumed to have a known distribution, e. g. normal.

If random parameters are not correlated then $\boldsymbol{\Gamma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$ is a diagonal matrix. To allow for correlated parameters, $\boldsymbol{\Gamma}$ is specified as a lower triangular matrix so that the variance-covariance matrix of the random coefficients becomes $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Sigma}$. Non-random parameters in the model can be easily incorporated in this formulation by specifying the corresponding rows in $\mathbf{D}$ and $\boldsymbol{\Gamma}$ to be zero. The conditional choice probability that individual i chooses alternative j, conditional on the realization of $\boldsymbol{\eta}_i$, is

$$P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) = \frac{\exp\left(\theta_j + \boldsymbol{\beta}_i' \tilde{\mathbf{X}}_{ij}\right)}{1 + \exp\left(\theta_i + \boldsymbol{\beta}_i' \tilde{\mathbf{X}}_{ij}\right)} \,, \tag{34}$$

where $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and $\boldsymbol{\eta}_i$ has distribution G with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{I}$.

Unconditional choice probability P_{ij} for alternative j is given by

$$P_{ij} = \int_{\boldsymbol{\eta}_i} P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) \mathrm{d}F_{\boldsymbol{\eta}}(\boldsymbol{\eta}_i) \,, \tag{35}$$

where $F_{\boldsymbol{\eta}}(.)$ is the joint c.d.f. of $\boldsymbol{\eta}_i$. The choice probability can be interpreted as a weighted average of logit probabilities with weights given by the mixing c.d.f. $F_{\boldsymbol{\eta}}(.)$. Following (10), the log-likelihood for θ is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \log P_{ij} \,. \tag{36}$$

The unconditional choice probability P_{nj} involves an integral over the mixing distribution, but the log-likelihood function does not generally have a closed form. Hence one cannot differentiate the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and solve the first order conditions in order to obtain the parameter estimates. Instead, one estimates the choice probability P_{ij} through simulation and then maximize the resulting simulated maximum likelihood (SML) with respect to the parameter vector.

Train [89] shows that this mixed logit framework leads to a tractable, unbiased and smooth simulator for the choice probability defined by:

$$\hat{P}_{ij} = \hat{P}(j|\mathbf{X}_i, \boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta}) \,, \tag{37}$$

where $\boldsymbol{\beta}_i^s = \mathbf{b} + \mathbf{D}\boldsymbol{\omega}_i + \boldsymbol{\Gamma}\eta_i^s$ and $\boldsymbol{\eta}_i^s$ is the sth ($s = 1, 2, \ldots, S$) draw from the joint distribution of $\boldsymbol{\eta}_i^s$, i. e., from $f(\boldsymbol{\eta}_i)$.

The log-likelihood function can be approximated by maximum simulated log-likelihood (MSL) given by

$$\begin{aligned} S\mathcal{L}(\boldsymbol{\theta}_\beta) &= \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \log \hat{P}_{ij} \\ &= \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \log \left[\frac{1}{S} \sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta}) \right] \,. \end{aligned} \tag{38}$$

Note that although $\hat{P}_j$ is unbiased for P_j, $\ln(\hat{P}_j)$ is not unbiased for $\ln(P_j)$, therefore the simulator generates some bias. To avoid bias, the simulation approximation should be improved. That means one must choose S to be sufficiently large. How large is "sufficiently large"? There is no fixed answer. But a result due to Gourieroux and Monfort [28] states indicates that the number should increase with the sample size N. Specifically, if the number of simulations, S, increases faster than the square root of the number of observations, this bias disappears in large samples. More pragmatically, the user should check that the results do not change much if S is increased.

To simulate the choice probability P_{ij}, one generally requires a large number of pseudo-random draws from the mixing distribution so that resulting simulation errors in the parameter estimates are kept at a reasonable level. Fortunately, advances in simulation methodology, such as the use of quasi-random numbers, in place of pseudo-random numbers, makes this feasible; see Train [89].

The preceding examples illustrate the point that accommodating heterogeneity in a flexible manner comes at a considerable computational cost. In many cases they lead to simulation-assisted estimation, this being an area of microeconometrics that has developed mainly since the 1990s.

Nonrepresentative Samples

Microeconometric methods often invoke the assumption that analysis is based on simple random samples (SRS). This assumption is hardly ever literally true for survey data. More commonly a household survey may first stratifies the population geographically into subgroups and applies differing sampling rates for different subgroups. An important strand in microeconometrics addresses issues of estimation and inference when the i.i.d. assumption no longer applies because the data are obtained from stratified and/or weighted samples. Stratified sampling methods also lead to dependence or clustering of cross section and panel observations. Clusters may have spatial, geographical, or economic dimension. In these cases the usual methods of establishing distribution of estimators based on the SRS assumption need to be adapted.

Stratified Samples For specificity it is helpful to mention some common survey stratification schemes. Table 1 based on Imbens and Lancaster [46] and Cameron and Trivedi [13], provides a summary.

Econometricians have paid special attention to endogenous stratification because this often leads to inconsistency of some standard estimation procedures such as ML; see Manski and Lerman [64], Cosslett [15], Manski and McFadden [65]. One example is choice-based sampling for binary or multinomial data where samples are chosen based on the discrete outcome y. For example, if choice is between travel to work by bus or travel by car we may over-sample bus riders if relatively few people commute by bus. A related example is count data on number of visits collected by on-site sampling of users, such as sampling at recreational sites or shopping centers or doctors offices. Then data are truncated, since those with $y = 0$ are not sampled, and additionally high frequency visitors are over-sampled.

Microeconometrics, Table 1
Alternative sample stratification schemes

Stratification Scheme	Description
Simple random	One strata covers entire sample space.
Pure exogenous	Stratify on regressors only, not on dependent variable.
Pure endogenous	Stratify on dependent variable only, not on regressors.
Augmented sample	Random sample augmented by extra observations from part of the sample space.
Partitioned	Sample space split into mutually exclusive strata that fill the entire sample space.

Endogenously stratified sampling leads to the density in the sample differing from that in the population (Cameron and Trivedi, pp. 822–827 in [13]). If the sample and population strata probabilities are known, then the standard ML and GMM estimation can be adapted to reflect the divergence. Typically this leads to weighted ML or weighted GMM estimation [46].

Clustered and Dependent Samples Survey data are usually dependent. This may reflect a feature of the survey sampling methodology, such as interviewing several households in a neighborhood. Consequently, the data may be correlated within cluster due to presence of a common unobserved cluster-specific term. Potentially, such dependence could also arise with SRS.

There are several different methods for controlling for dependence on unobservables within cluster. If the within cluster unobservables are uncorrelated with regressors then only the variances of the regression parameters need to be adjusted. This leads to cluster-correction-of-variances methods that are now well-embedded in popular software packages such as Stata. If, instead, the within cluster unobservables are correlated with regressors then the regression parameters are inconsistent and fixed effects type methods are called for. The issues and available methods closely parallel those for fixed and random effects panels models. Further, methods may also vary according to whether there are many small clusters or few large clusters. Examples and additional detail are given in Cameron and Trivedi [13].

An important new topic concerns dependence in cross section and panel data samples between independently obtained measures. Several alternative models are available to motivate such dependence. Social interactions [21] between individuals or households, and spatial dependence [7] where the observational unit is region, such as state, and observations in regions close to each other are likely to be interdependent, are examples. Models of social interaction analyze interdependence between individual choices (e. g. teenage smoking behavior) due to, for example, peer group effects. Such dependence violates the commonly deployed i.i.d. assumption, and in some cases the endogeneity assumption. Lee [57] and Andrews [4] examine the econometric implications of such dependence.

Major Insights

A major role of microeconometrics is inform public policy. But public policy issues arise not only in the context of existing policies whose effectiveness needs evaluation but also for new policies that have never been tried and old policies that are candidates for adoption in new economic environments. No single approach to microeconometrics is appropriate for all these policy settings. All policy evaluation involves comparison with counterfactuals. The complexity associated with generating counterfactuals varies according to the type of policy under consideration as well as the type of data on which models are based. A deeper understanding of this fundamental insight is a major contribution of modern microeconometrics.

A second major insight is the inherent difficulty of making causal inferences in econometrics. Many different modeling strategies are employed to overcome these challenges. At one end of the spectrum are highly structured models that make heavy use of behavioral, distributional and functional form assumptions. Such models address more detailed questions and provide, conditional on the framework, more detailed estimates of the policy impact. At the other end of the spectrum are methods that minimize on assumptions and aim to provide informative bounds for measures of policy impact. While the literature remains unsettled on the relative merits and feasibility of these approaches, the trend in microeconometrics is towards fewer and less restrictive assumptions.

There is now a greater recognition of the challenges associated with analyzes of large complex data sets generated by traditional sample surveys as well as other automated and administrative methods. In so far as such challenges are computational, advances in computer hardware and software technologies have made a major contribution to their solution.

Bibliography

Primary Literature

1. Albert JH, Chib S (1993) Bayesian Analysis of Binary and Polychotomous Response Data. J Am Stat Assoc 88:669–679
2. Allen RGD, Bowley AL (1935) Family Expenditure. PS King and Son, London
3. Amemiya T (1985) Advanced Econometrics. Harvard University Press, Cambridge (Mass)
4. Andrews DWK (2005) Cross-section Regression with Common Shocks. Econometrica 73:1551–1585
5. Angrist JD, Krueger AB (2000) Empirical Strategies in Labor Economics. In: Ashenfelter OC, Card DE (eds) Handbook of Labor Economics, vol 3A. North-Holland, Amsterdam, pp 1277–1397
6. Angrist JD, Imbens G, Rubin D (1996) Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc 91:444–455
7. Anselin L (2001) Spatial Econometrics. In: Baltagi BH (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, pp 310–330
8. Baker M, Melino A (2000) Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study. J Econ 96:357–393
9. Blundell RW, Powell JL (2001) Endogeneity in Semiparametric Binary Response Models. Rev Econ Stud 71:581–913
10. Blundell RW, Smith RJ (1989) Estimation in a Class of Limited Dependent Variable Models. Rev Econ Stud 56:37–58
11. Bunch D (1991) Estimability in the Multinomial Probit Model. Transp Res Methodol 25B(1):1–12
12. Cameron AC, Trivedi PK (1998) Regression Analysis for Count Data. Econometric Society Monograph No. 30, Cambridge University Press, Cambridge
13. Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge
14. Chesher A (2005) Nonparametric identification under discrete variation. Econometrica 73(5):1525–1550
15. Cosslett SR (1981) Maximum Likelihood Estimator for Choice-Based Samples. Econometrica 49:1289–1316
16. Deb P, Trivedi PK (1997) Demand for Medical Care by the Elderly: A Finite Mixture Approach. J Appl Econ 12:313–326
17. Deb P, Trivedi PK (2002) The Structure of Demand for Health Care: Latent Class versus Two-part Models. J Health Econ 21:601–625
18. Deb P, Trivedi PK (2006) Specification and Simulated Likelihood Estimation of a Non-normal Treatment-Outcome Model with Selection: Application to Health Care Utilization. Econ J 9:307–331

19. Diggle PJ, Heagerty P, Liang KY, Zeger SL (1994, 2002) Analysis of Longitudinal Data, 1st and 2nd editions. Oxford University Press, Oxford
20. Dubin J, McFadden D (1984) An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. Econometrica 55:345–362
21. Durlauf S, Cohen-Cole E (2004) Social Interactions Models. In: Lempf-Leonard K (ed) Encyclopedia of Social Measurement. Academic Press, New York
22. Eckstein Z, Wolpin K (1989) The Specification and Estimation of of Dynamic Stochastic Discrete Choice Models: A Survey. J Hum Resour 24:562–598
23. Engel E (1857) Die Produktions- und Consumptionsverhältnisse des Königreichs Sachsen, Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministerium des Innern, 22 November 1857. Reprinted in 1895 as appendix to E. Engel, Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt. Bull Inst Int Stat 9:1–124
24. Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, New York
25. Geweke J (1992) Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments (with discussion). In: Bernardo J, Berger J, Dawid AP, Smith AFM, (eds) Bayesian Statistics, vol 4. Oxford University Press, Oxford, pp 169–193
26. Greene WH (2004) The Behavior of the Fixed Effects Estimator in Nonlinear Models. Econ J 7(1):98–119
27. Greene WH (2004) Fixed Effects and the Incidental Parameters Problem in the Tobit Model. Econ Rev 23(2):125–148
28. Gouriéroux C, Monfort A (1996) Simulation Based Econometrics Methods. Oxford University Press, New York
29. Hajivassiliou V, McFadden D, Ruud P (1996) Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. J Econ 72:85–134
30. Heckman JJ (1974) Shadow Prices, Market wages, and Labor Supply. Econometrica 42:679–94
31. Heckman JJ (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. Ann Econ Soc Meas 5:475–492
32. Heckman JJ (1978) Dummy Endogenous Variables in a Simultaneous Equations System. Econometrica 46:931–960
33. Heckman JJ (1979) Sample Selection as a Specification Error. Econometrica 47:153–61
34. Heckman JJ (2000) Causal Parameters and Policy Analysis in Economics: A Twentieth Century Perspective. Quart J Econ 115:45–98
35. Heckman JJ (2001) Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. J Political Econ 109(4):673–748
36. Heckman JJ, Robb R (1985) Alternative Methods for Evaluating the Impact of Interventions. In: Heckman JJ, Singer B (eds) Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge, UK
37. Heckman JJ, Singer B (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration Data. Econometrica, 52:271–320
38. Heckman JJ, Vytlacil E (2001) Local instrumental variables. In: Hsiao C, Morimue K, Powell JL (eds) Nonlinear Statistical Modeling. In: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in the Honor of Takeshi Amemiya. Cambridge University Press, New York, pp 1–46
39. Heckman JJ, Vytlacil E (2005) Structural Equations, Treatment Effects, and Econometric Policy Evaluation. Econometrica 73(3):669–738
40. Holland PW (1986) Statistics of Causal Inference. J Am Stat Assoc 81:945–60
41. Horowitz J (1998) Semiparametric Methods in Econometrics. Springer, New York
42. Hotz V, Miller R (1993) Conditional Choice Probabilities and the Estimation of Dynamic Models. Rev Econ Stud 60:497–529
43. Houthakker HS (1957) An International Comparison of Household Expenditure Patterns. Econometrica 25:532–552
44. Hsiao C (1986, 2003) Analysis of Panel Data, 1st and 2nd edn. Cambridge University Press, Cambridge
45. Imbens GW, Angrist J (1994) Identification and Estimation of Local Average Treatment Effect. Econometrica 62:467–475
46. Imbens GW, Lancaster T (1996) Efficient Estimation and Stratified Sampling. J Econ 74:289–318
47. Keane MP (1992) A Note on Identification in the Multinomial Probit Model. J Bus Econ Stat 10:193–200
48. Keane MP (1994) A Computationally Practical Simulation Estimator for Panel Data. Econometrica 62:95–116
49. Keane MP (2006) Structural vs. Atheoretical Approaches to Econometrics. Unpublished paper
50. Keane M, Wolpin K (1994) The Solutions and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpretation: Monte Carlo Evidence. Rev Econ Stat 76:648–672
51. Koenker R (2005) Quantile Regression. Cambridge University Press, New York
52. Koenker R, Bassett G (1978) Regression Quantiles. Econometrica 46:33–50
53. Koop G, Poirier D, Tobias JL (2007) Bayesian Econometric Methods. Cambridge University Press, Cambridge
54. Lancaster T (2000) The Incidental Parameter Problem since 1948. J Econ 95:391–413
55. Lancaster T, Imbens GW (1996) Case-Control with Contaminated Controls. J Econ 71:145–160
56. Lee LF (2001) Self-selection. In: Baltagi B (ed) A Companion to Theoretical Econometrics. Blackwell Publishers, Oxford, pp 381–409
57. Lee LF (2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models. Econometrica 72:1899–1926
58. Lee MJ (2002) Micro-Econometrics for Policy, Program and Treatment Effects. Oxford University Press, Oxford
59. Luce D (1959) Individual Choice Behavior. Wiley, New York
60. Ma L, Koenker R (2006) Quantile regression methods for recursive structural equation models. J Econ 134:471–506
61. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3:205–228
62. Manski CF (1995) Identification Problems in the Social Sciences. Harvard University Press, Cambridge
63. Manski CF (2003) Partial Identification of Probability Distributions. Springer, New York
64. Manski CF, Lerman SR (1977) The Estimation of Choice Probabilities from Choice-Based Samples. Econometrica 45:1977–1988

65. Manski CF, McFadden D (1981) Alternative Estimators and Sample Design for Discrete Choice Analysis. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, pp 2–50
66. Manski CF, McFadden D (eds) (1981) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge
67. McCulloch R, Rossi P (2000) Bayesian analysis of the multinomial probit model. In: Mariano R, Schuermann T, Weeks M (eds) Simulation-Based Inference in Econometrics. Cambridge University Press, New York
68. McFadden D (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P (ed) Frontiers of Econometrics. Academic Press, New York
69. McFadden D (1976) Quantal Choice Analysis: A Survey. Ann Econ Soc Meas 5(4):363–390
70. McFadden D (1978) Modelling the Choice of Residential Location. In: Karlquist A et al (eds) Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, New York
71. McFadden D (1981) Econometric Models of Probabilistic Choice. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Economic Applications. MIT Press, Cambridge, pp 198–272
72. McFadden D (1984) Econometric Analysis of Qualitative Response Models. In: Griliches Z, Intriligator M (eds) Handbook of Econometrics, vol 2. North Holland, Amsterdam
73. McFadden D (2001) Economic Choices. Am Econ Rev 91(3):351–78
74. McFadden D, Ruud PA (1994) Estimation by Simulation. Rev Econ Stat 76:591–608
75. McFadden D, Train K (2000) Mixed MNL Models of Discrete Response. J Appl Econ 15:447–470
76. Maddala GS (1983) Limited Dependent and Qualitative Variables in Economterics. Cambridge University Press, Cambridge
77. Marschak J, Andrews WH (1944) Random Simultaneous Equations and the Theory of Production. Econometrica 12:143–205
78. Miller R (1997) Estimating Models of Dynamic Optimization with Microeconomic Data. In: Pesaran HH, Schmidt P (eds) Handbook of Applied Econometrics, vol II. Blackwell Publishers, Oxford
79. Prais SJ, Houthakker HS (1955) Analysis of Family Budgets. Cambridge University Press, Cambridge
80. Pudney S (1989) Modeling Individual Choice: The Econometrics of Corners. Basil, Kinks and Holes. Blackwell, New York
81. Rust J (1987) Optimal replacement of gmc bus engines: An empirical model of Harold Zurcher. Econometrica 55:993–1033
82. Rust J (1994) Estimation of dynamic structural models, problems and prospects: Discrete decision processes. In: Sims C (ed) Advances in Econometrics: Sixth World Congress, vol II. Cambridge University Press, New York, pp 5–33
83. Rust J (1994) Structural Estimation of Markov Decision Processes. In: Engle RF, McFadden D (eds) Handbook of Econometrics, vol 4. North-Holland, Amsterdam, pp 3081–3143
84. Rust J (1997) Using Randomization to Break the Curse of Dimensionality. Econometrica 65:487–516
85. Rust J, Phelan C (1997) How Social Security and Medicare Affect Treatment Behavior in a World of Incomplete Markets. Econometrica 65(4):781–842
86. Stone R (1953) The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, vol 1. Cambridge University Press, Cambridge, pp 1920–1938
87. Thurstone L (1927) A Law of Comparative Judgment. Psychol Rev 34:273–286
88. Tobin J (1958) Estimation of Relationships for Limited Dependent Variables. Econometrica 26:24–36
89. Train KE (2003) Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge
90. Walker J, Ben-Akiva M (2002) Generalized random utility model. Math Soc Sci 43:303–343

Books and Reviews

Arellano M (2003) Panel Data Econometrics. Oxford University Press, Oxord, UK

Blundell R, Powell JL (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, vol II. Cambridge University Press, Cambridge

Deaton A (1997) The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins, Baltimore

Greene WH (2007) Econometric Analysis, 6th edn. Macmillan, New York

Hensher DA, Rose J, Greene W (2005) Applied Choice Analysis: A Primer. Cambridge University Press, New York

Lancaster T (1990) The Econometric Analysis of Transitional Data. Cambridge University Press, New York

Lee MJ (2002) Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables. Academic Press, San Diego

Louviere J, Hensher D, Swait J (2000) Stated Choice Methods: Analysis and Applications. Cambridge University Press, New York

Wooldridge JM (2002) Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge

Yatchew A (2003) Semiparametric Regression for the Applied Econometrician. Cambridge University Press, Cambridge

Nonparametric Tests for Independence

Cees Diks
University of Amsterdam, Amsterdam, The Netherlands

Article Outline

Glossary

Hypothesis A hypothesis is a statement concerning the (joint) distribution underlying the observed data.

Nonparametric test In contrast to a parametric test, a nonparametric test does not presume a particular parametric structure concerning the data generating process.

Serial dependence Statistical dependence among time series observations.

Time series A sequence of observed values of some variable over time, such as a historical temperature record, a sequence of closing prices of a stock, etc.

Definition of the Subject

One of the central goals of data analysis is to measure and model the statistical dependence among random variables. Not surprisingly, therefore, the question whether two or more random variables are statistically independent can be encountered in a wide range of contexts. Although this article will focus on tests for independence among time series data, its relevance is not limited to the time series context only. In fact many of the dependence measures discussed could be utilized for testing independence between random variables in other statistical settings (e. g. cross-sectional dependence in spatial statistics).

When working with time series data that are noisy by nature, such as financial returns data, testing for serial independence is often a preliminary step carried out before modeling the data generating process or implementing a prediction algorithm for future observations. A straightforward application in finance consists of testing the random walk hypothesis by checking whether increments of, for instance, log prices or exchange rates, are independent and identically distributed [8,12,80]. Another important application consists of checking for remaining dependence structure among the residuals of an estimated time series model.

Introduction

Throughout this article it will be assumed that $\{X_t\}$, $t \in \mathbb{Z}$, represents a strictly stationary time series process, and tests for serial independence are to be based on an observed finite sequence $\{X_t\}_{t=1}^n$. Unless stated otherwise, it will be assumed that the observations X_t take values in the real line $\mathbb{R}$. Admittedly, this is a limitation, since there are also time series processes that do not take values in the real line. For instance, the natural space in which wind direction data take values is the space of planar angles, which are naturally represented by the interval $[0, 2\pi]$ with the endpoints identified. However, most of the tests developed to date are designed for the real-valued case. The problem under consideration is that of testing the null hypothesis that the time series process $\{X_t\}$ consists of independent, identically distributed (i.i.d.) random variables. In practice this is tested by looking for dependence among of m consecutive observations $X_{t-m+1}, \ldots, X_t$ for a finite value $m \geq 2$.

Traditionally, tests for serial independence have focused on detecting serial dependence structure in stationary time series data by estimating the autocorrelation function (acf), $\rho_k = \mathrm{Cov}(X_{t-k}, X_t)/\mathrm{Var}(X_t)$, or the normalized spectral density, which is one-to-one related to the acf by Fourier transformation:

$$h(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \rho_k \mathrm{e}^{-i\omega k}$$

$$\text{and} \quad \rho_k = \int_{-\pi}^{\pi} h(\omega) \mathrm{e}^{ik\omega} \mathrm{d}\omega \,.$$

Because the acf is real and symmetric ($\rho_k = \rho_{-k}$) the normalized spectral density is real and symmetric also. Since the acf and the normalized spectral density are related by an invertible transformation, they carry the same information regarding the dependence of a process. For i.i.d. processes with finite variance, $\rho_k = 0$ for $k \geq 1$ under the null hypothesis. The spectral density is flat (equal to 1 for all ω) in that case.

Tests based on the acf date back to Von Neumann [83]. Motivated by the aim to test for serial independence against the presence of trends, he introduced the ratio of the mean square first difference to the sample variance,

$$S_n := \frac{\frac{1}{n-1}\sum_{t=2}^{n}(X_t - X_{t-1})^2}{\frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2} ,$$

which may be considered a rescaled (and shifted) estimator of ρ_1. Von Neumann studied the distributional properties of this statistic in detail under the assumption of normality. Durbin and Watson [38,39] used an analogue of Von Neumann's ratio to check for first order autocorrelation among the error terms $\{\varepsilon_t\}_{t=1}^{n}$ in a linear regression model, based on observed residuals $\{\widehat{\varepsilon}_t\}_{t=1}^{n}$. As for the original statistic of Von Neumann, the null distribution (which is no longer unique in this case, but depends on the parameters of the data generating process) has been studied in detail for the normal case [40,56,101]. For the class of autoregressive integrated moving average (ARIMA) processes, acf-based tests for residual autocorrelation beyond lag 1 were proposed by Box and Jenkins [13] and Ljung and Box [79], and for autocorrelation in squared residuals by McLeod and Li [82]. Beran [9] proposed adapted tests for serial independence for processes with long-range dependence.

Although the autocovariance structure of a time series process fully characterizes the dependence structure within classes of linear Gaussian random processes, tests based solely on the acf may clearly fail to be consistent against dependence that does not show up in the acf. It is not hard to construct examples of processes for which this is the case. For instance, the bilinear process

$$X_t = aX_{t-2}\varepsilon_{t-1} + \varepsilon_t , \quad (|a| < 1)$$

where $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables, clearly exhibits dependence, but has no autocorrelation structure beyond lag zero. Other examples include the ARCH(1) process [42],

$$X_t = \sqrt{h_t}\varepsilon_t , \quad h_t = c + \theta X_{t-1}^2 , \quad (c > 0, 0 \le \theta < 1)$$

and the GARCH(1,1) process [11],

$$X_t = \sqrt{h_t}\varepsilon_t , \quad h_t = c + \alpha h_{t-1} + \beta X_{t-1}^2 ,$$
$$(c > 0, \alpha, \beta > 0, \alpha + \beta < 1)$$

which have become popular for modeling financial returns data.

The desire to avoid specific assumptions about the process under the null hypothesis or under the possible alternatives motivates a nonparametric statistical approach to the problem of testing for serial independence. One possibility is to develop rank-based tests for serial independence against particular types of structure such autoregressive moving average (ARMA) structure. Compared to the linear Gaussian paradigm, this approach explicitly drops the assumption that the marginal distribution is normal, which in a natural way leads to tests formulated in terms of ranks. This has the advantage that the tests are distribution free (the null distributions of test statistics do not depend on the actual marginal distribution). The developments around invariant tests for serial independence are reviewed briefly in Sect. "Invariant Tests".

Another nonparametric approach consists of using nonparametric estimators of divergence measures between distributions to construct tests against unspecified alternatives. The idea is to measure the discrepancy between the joint distribution of $(X_{t-m+1}, \ldots, X_t)$ and the product of marginals with a measure of divergence between multivariate probability measures. This typically involves estimating a suitable measure of dependence, and determining the statistical significance of the observed value of the statistic for the sample at hand. In Sect. "Tests Based on Divergence Measures" several tests for serial independence based on divergence measures are reviewed. For further details on some of the earlier methods, the interested reader is referred to the overview by Tjøstheim [103].

Section "Tests Based on Other Measures of Dependence" describes tests for independence based on some other measures of serial dependence in the observed time series, such as partial sums of observations and the bispectrum.

For particular statistics of interest critical values can be obtained in different ways. The traditional way is to use critical values based on asymptotic theory, which is concerned with the large sample limiting distributions of test statistics. With the increasing computer power that became available to most researchers in the recent decades, it has become more and more popular to obtain critical values of test statistics by resampling and other computer simulation techniques. I will discuss the advantages and disadvantages of several of these numerical procedures in Sect. "Bootstrap and Permutation Tests".

Note that pairs of delay vectors such as (X_{t-1}, X_t) and $(X_{s-1}, X_s)'$ for $s \neq t$ may have elements in common, and hence are not independent even under the null; a fact which has to be taken into account when critical values of test statistics are determined. Tests for independence among m random variables, $(Y_1, \ldots, Y_m)$, say, based on a random sample therefore typically need to be adapted for applications in a time series setting. For most asymptotic results that depend on the distribution of the test statistic under the alternative (such as consistency) additional assumptions are required on the rate of decay of the dependence in the data, known as mixing conditions [14].

Notation

Let $\boldsymbol{X}_t^m$ be short-hand notation for the delay vector $(X_{t-m+1}, \ldots, X_t)$, $m \geq 3$. For the case $m = 2$, $\boldsymbol{X}_t^m$ refers to a bivariate vector (X_{t-k}, X_t), $k \geq 1$, where the value of k will be clear from the context. Under the assumption that X_t takes values in the real line $\mathbb{R}$ (or a subset thereof) one may state the null hypothesis in terms of the joint and marginal cumulative distribution functions (CDFs):

$$H_0: \quad F_m(\boldsymbol{x}) = F_1(x_1) \times \cdots \times F_1(x_m) ,$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)'$, and $F_m(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_m \leq x_m)$ is the joint cumulative distribution function (CDF) of $\boldsymbol{X}_t^m$, and $F_1(x) = P(X \leq x)$ the marginal CDF of $\{X_t\}$. If $\boldsymbol{X}_t^m$ is a continuous random variable, one can denote its probability density function by $f_m(\boldsymbol{x})$, and the independence of the elements of $\boldsymbol{X}_t^m$ can be written as

$$H_0: \quad f_m(\boldsymbol{x}) = f_1(x_1) \times \cdots \times f_1(x_m) ,$$

where $f_1(x)$ is the marginal probability density function of $\{X_t\}$. For convenience I will drop the subscript m in $f_m(\boldsymbol{x})$, and introduce $g(\boldsymbol{x}) := f_1(x_1) \times \cdots \times f_1(x_m)$, so that the null hypothesis can be rephrased simply as

$$H_0: \quad f(\boldsymbol{x}) = g(\boldsymbol{x}) .$$

Some Practical Considerations

Which of the tests described below should one choose for practical applications? The alternative against which the null hypothesis is to be tested is any deviation from the above factorizations for some $m \geq 2$. Ideally, one would like a nonparametric test to have large power against all types of dependence. However, since no uniformly most powerful test against all possible alternatives exists, among the tests proposed in the literature one typically finds that some perform better against certain alternatives and some against others, and it is often hard to identify exactly why. Although power against the alternative at hand is obviously important in applications, usually these alternatives are not known in a simple parametric form. This is precisely what motivated many of the recently developed tests tests for independence; they are designed to have power against large classes of alternatives. When a practitioner has to choose among the large number of the omnibus tests available, besides power also some other properties can be taken into consideration. For instance, some tests are invariant (immune to invertible transformations of the data, see Sect. "Invariant Tests") while others, such as those based on true divergence measures discussed in Sect. "Tests Based on Divergence Measures", are consistent against any fixed alternative, which means that they will asymptotically (with increasing sample size) detect any given alternative with probability one.

Although invariance is a pleasant property, because it allows one to tabulate the null distribution, I would generally rank consistency against any fixed alternative as more important, since invariance can usually be achieved easily by a simple trick, such as transforming the data to ranks before applying any given independence test. At the same time I should add that if one is willing to settle for power against particular classes of alternatives only, it is sometimes possible to construct an ideal hybrid between invariance and consistency in the form of an optimal invariant test. An example is the optimal rank test of Benghabrit and Hallin [7] discussed in the next section.

A clear disadvantage of omnibus tests is that after a rejection of the null hypothesis it leaves the practitioner with the problem of having to identify the type of dependence separately. If one is confident enough to assume (or lucky enough to know) a specific parametric form for the data generating process it is arguably more efficient to rely on traditional parametric methods. However, I think that in most cases the potential efficiency gains are not worth the risk of biased test results due to a misspecification of an unknown type.

Invariant Tests

When developing nonparametric tests for serial independence it is typically assumed that the marginal distribution of the observed time series process is unknown. Because in general the distribution of the test statistic will depend on this unknown distribution, the latter plays the role of an infinite dimensional nuisance parameter. There are various ways of dealing with this problem, such as focusing on particular classes of test statistics and appealing to asymptotic theory, or using bootstrap or permutation techniques. These methods are rather common and are used in most of the tests discussed in the subsequent sections. This section is concerned with a more direct way to deal with the nuisance parameter problem. The main idea is to focus on dependence measures that are invariant under one-to-one transformations of the space in which X_t takes values (so-called static transformations $X_t' = \phi(X_t)$, where ϕ is a strictly monotonous map from $\mathbb{R}$ to itself). This naturally leads to the study of statistics based on ranks.

Rank Tests

Various analogues of the correlation coefficient have been proposed based on ranks. For the pairs $\{(X_t, Y_t)\}$, Spearman's rank correlation [97] is the sample correlation of R_t

and S_t, the ranks of X_t and Y_t among the observed X's and the Y's, respectively. In a univariate time series context one can easily define a serial version of this rank correlation (e. g. the sample autocorrelation function of the sequence of ranks $\{R_t\}$ of the X's). Kendall's tau [73] for pairs $\{(X_t, Y_t)\}$ is another rank-based measure of dependence, quantifying the concordance of the signs of $X_i - X_j$ and $Y_i - Y_j$. The serial version of tau can be defined as

$$\tau_k = \binom{n-k}{2} \sum_{i=1}^{n-k} \sum_{j=1}^{i-1} \operatorname{sgn}(X_i - X_j) \operatorname{sgn}(X_{i+k} - X_{j+k}) \,.$$

The multivariate versions of these concordance orderings have been described by Joe [69]. Genest et al. [45] considered tests for serial independence, building on asymptotic results derived by Ferguson et al. [43] for a serialized version of Kendall's tau in a time series setting.

Many other rank-based tests for independence have been developed meanwhile. The earlier work in this direction is covered in the review paper by Dufour [36]. Later work includes that by Bartels [6], who developed a rank-based version of Von Neumann's statistic, Hallin et al. [54] who proposed rank-based tests for serial independence against ARMA structure, and Hallin and Mélard [55] who study the finite sample behavior and robustness against outliers of their proposed procedures. Kallenberg and Ledwina [72] developed a nonparametric test for the dependence of two variables by testing for dependence of the joint distribution of ranks in a parametric model for the rank dependence.

Optimal rank tests are tests based on ranks that have maximal power. Naturally such a test depends on the alternative against which the power is designed to be large. For instance, Benghabrit and Hallin [7] derived an optimal rank test for serial independence against superdiagonal bilinear dependence.

As a way to deal with the problem that the marginal distribution is unknown under the null hypothesis (the nuisance parameter problem) Genest, Ghoudi and Rémillard [48] consider rank-based versions of the BDS test statistic (see Sect. "Correlation Integrals"), as well as several other rank-based statistics.

Empirical Copulae

As noted by Genest and Rémillard [46], a rank test for serial independence can alternatively be considered a test based on the empirical copula. The reason is that the empirical copula determines the sequence of ranks and vice versa. I therefore briefly review the notion of a copula.

If $\boldsymbol{X}_t^m$ is a continuous random variable, its copula C is defined as

$$F(x_1, \ldots, x_m) = C(F(x_1), \ldots, F(x_m)) \,,$$

where $F(x)$ denotes the marginal CDF. Note that $C(u_1, \ldots, u_m)$ is defined on the unit (hyper-)cube $[0,1]^m$ (unit square for $m = 2$), and has the properties of a CDF of some distribution on that space. This allows one to define the associated copula density on the unit cube as

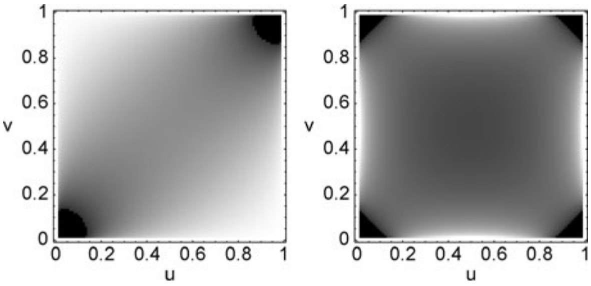
$$c(u_1, \ldots, u_m) = \frac{\partial^m}{\partial u_1 \cdots \partial u_m} C(u_1, \ldots, u_m) \,.$$

The copula density c is obtained by taking partial derivatives with respect to each of the X_i's:

$$f(x_1, \ldots, x_m) = c(F(x_1), \ldots, F(x_m)) \times f(x_1) \times \cdots \times f(x_m).$$

The null hypothesis of serial independence states $f(\boldsymbol{x}) = g(\boldsymbol{x}) = f(x_1) \times \cdots \times f(x_m)$, which is equivalent to $c(u_1, \ldots, u_m) = 1$. This shows that the factorization of the joint distribution in the product of marginals really is a property of the copula. In this sense the copula can be viewed as containing all relevant information regarding the dependence structure of $\boldsymbol{X}_t^m$. Figure 1 shows the Gaussian copula for a bivariate distribution with correlation coefficient 0.5 and the local ARCH(1) copula (essentially a rescaled version of the copula obtained for an infinitesimal positive ARCH parameter).

The empirical copula obtained from time series data is the empirical distribution of $(\widehat{U}_{t-m+1}, \ldots, \widehat{U}_t)$ where $\widehat{U}_t$ is the normalized rank of X_t, defined as $\widehat{U}_t = \#\{X_s \leq X_t\}/n$. Assuming that ties (identical values of X_t and X_s for $t \neq s$) are absent, each rank only occurs once, and hence the empirical copula is one-to-one connected to the sequence of ranks. This shows that test based on ranks can be considered as tests based on the empirical copula and vice versa. It also shows that the concept of an optimal



Nonparametric Tests for Independence, Figure 1
The Gaussian copula density for $\rho = 0.5$ (*left*) and local ARCH(1) copula density (*right*)

rank test against a particular copula alternative is meaningful.

The connection between sequences of ranks and empirical copulae makes it rather intuitive to design tests that have high power against serial dependence described by particular (families of) copulae. Genest and Verret [47] consider rank-based tests for independence of two random variables that are locally most powerful against a number of parametric copulae. Scaillet [92] used the copula representation to test for serial independence against positive quadrant dependence between two random variables, which holds if $P[X \leq x, Y \leq Y] \geq P[X \leq x]P[Y \leq y]$, or equivalently $P[X > x, Y > y] \geq P[X > x]P[Y > y]$. In a similar vein Panchenko and I [33] derived a rank test for serial independence against the local ARCH(1) copula.

Tests Based on Divergence Measures

In this section I consider tests for serial independence based on various dependence measures. Typically the tests obtained with this approach are not invariant. However, critical values of test statistics can still be obtained using asymptotic theory or bootstrap methods (see Sect. "Bootstrap and Permutation Tests" for more details), and the tests are consistent against a wide range of alternatives.

Many popular dependence measures are based on divergences between the m-dimensional density f and its counterpart under the null hypothesis, g. Divergences are functionals of pairs of densities, which, like distances, are equal to zero whenever $f(\boldsymbol{x}) = g(\boldsymbol{x})$ and strictly positive otherwise. To qualify as a true distance notion between distributions a divergence measure must also be symmetric and satisfy the triangle inequality. Not all divergence measures discussed below are true distances in this sense. This is no problem for testing purposes, but if one is interested in comparing distances with other distances (e. g. for cluster analysis) then the triangle inequality is essential [81]. In general, a divergence measure might serve just as well as a distance as a basis for constructing a test for serial independence.

Tests for serial independence can roughly be divided into two groups: tests against specific types of dependence and omnibus tests, with power against general types of dependence. For instance, the test of Von Neumann [83] mentioned above is sensitive to linear correlation between X_{t-1} and X_t, but is completely insensitive to some other types of dependence between these two variables. One of the great advantages of tests based on divergence measures is their omnibus nature. Typically these tests are consistent against any type of dependence among the m components of $\boldsymbol{X}_t^m$. Unfortunately, however, as noted in the Introduction, no uniformly most powerful test against serial independence exists, so different tests will be more powerful against different alternatives. Therefore, which test performs best in practice depends on the type of dependence structure present in the data.

Below I describe tests based on empirical distribution functions (empirical CDFs) as well as on densities. One can roughly state that tests based on the empirical CDFs are better suited to detecting large-scale deviations from the null distribution than small-scale deviations. In order for these tests to pick up deviations from independence there must be relatively large regions in $\mathbb{R}^m$ where the density of $\boldsymbol{X}_t^m$ is lower or higher than the hypothetical product density; the cumulative nature of the test statistics is relatively insensitive to small-scale deviations from the null distribution. If one wants to be able to detect subtle small-scale deviations between densities locally in the sample space, it seems more natural to use a test divergence measures based on density ratios or density differences, such as information theoretic divergences or correlation integrals. Note, however, that even among those tests performance may be hard to predict beforehand. For instance, in Subsect. "Information Theoretic Divergence Measures" I consider a family of closely related information theoretical divergence measures, but even within this family the relative powers of the tests depend strongly on the alternative at hand.

Empirical Distribution Functions

Empirical distribution functions have been used for studying the independence of random variables at least since Hoeffding [62], who proposed dependence measures based on the difference between the joint distribution function and the product of marginals, $F(\boldsymbol{x}) - G(\boldsymbol{x})$, where $G(\boldsymbol{x}) = \prod_{i=1}^m F_1(x_i)$ is the joint CDF under the null hypothesis.

There are various ways to define divergences in terms of distribution functions. A popular class of statistics is obtained by considering divergence measures of the type

$$d_w^2 = \int_{\mathbb{R}^m} (F(\boldsymbol{x}) - G(\boldsymbol{x}))^2 \, w(F(\boldsymbol{x})) \, \mathrm{d}F(\boldsymbol{x}) \,,$$

where $w(\cdot)$ is a positive weight function. For $w = 1$ this divergence is known as the Cramér–von Mises criterion, which has become a well-known criterion in univariate one- and two-sample problems. The Cramér–von Mises criterion suggests testing the independence of the elements of $\boldsymbol{X}_t^m$ based on its sample version

$$\tilde{n} d_n^2 = \sum_{i=m}^{n} \left(\widehat{F}(\boldsymbol{X}_i^m) - \widehat{G}(\boldsymbol{X}_i^m) \right)^2 \,,$$

where $\tilde{n} = n - m + 1$ is the number of m-dimensional delay vectors, $\widehat{F}$ is the empirical joint CDF and $\widehat{G}(\boldsymbol{x}) = \prod_{i=1}^{m} \widehat{F}_1(x_i)$ is the product of marginal empirical CDFs. This statistic, referred to as the Cramér–von Mises statistic, was proposed by Hoeffding [62] for testing independence in the bivariate case, based on a random sample of variables (X_i, Y_i) from a bivariate distribution (i. e. outside the time series scope). Since the statistic is invariant under one-to-one transformations of marginals, the tests based on it are automatically distribution-free. Although the null distribution for a random sample was known to converge in distribution to a mixture of scaled $\chi^2(1)$ random variables,

$$\tilde{n} d_n^2 \xrightarrow{d} \frac{1}{\pi^4} \sum_{j,k=1}^{\infty} \frac{1}{j^2 k^2} Z_{jk}^2 \,,$$

where the Z_{jk} are independent standard normal random variables, it was initially not available in a form suitable to practical applications. The distribution was tabulated eventually by Blum et al. [10], who also considered higher-variate versions of the test.

By generalizing results of Carlstein [23], Skaug and Tjøstheim [95] extended the test of Hoeffding, Blum, Kiefer and Rosenblatt to a time series context and derived the null distribution of the test statistic for first-order dependence ($m = 2$) under continuous as well as discrete marginals. In the continuous case it turns out that the first order test statistic has the same limiting null distribution as for a random sample from a bivariate distribution with independent marginals. Skaug and Tjøstheim [95] also showed that the statistic $nG_{K,n} = n \sum_{k=1}^{K} d_{k,n}^2$, where $d_{k,n}^2$ is the Cramér–von Mises statistic for (X_{t-k}, X_t), has a mixture of scaled $\chi^2(K)$ distributions as its limiting distribution. For high lags and moderate sample sizes Skaug and Tjøstheim [95] report that the asymptotic approximation to the finite sample null distribution is poor, and suggest the use of bootstrap methods to obtain critical values.

Delgado [29] considered the analogue test for higher-variate dependence in time series. In that case differences with the Hoeffding–Blum–Kiefer–Rosenblatt asymptotics arise due to the presence of dependence across the delay vectors constructed from the time series. Delgado and Mora [28] investigated the test for first order independence when applied to regression residuals, and found that the test statistic in that case has the same limiting null distribution as for serially independent data.

The Kolmogorov–Smirnov statistic

$$\sqrt{n} \sup_{\boldsymbol{x}} |\widehat{F}(\boldsymbol{x}) - \widehat{G}(\boldsymbol{x})| \,.$$

is another popular test statistic for comparing empirical cumulative distribution functions. Ghoudi et al. [49] developed asymptotic theory for this test statistic in rather general settings, which include the time series context. In their power simulations against several alternatives, including AR(1) and nonlinear moving average processes, the Cramér–von Mises statistic displayed a better overall performance than the Kolmogorov–Smirnov statistic.

This suggests that the Cramér–von Mises statistic might be a good choice for practical applications, provided that one wishes to compare empirical distribution functions. As noted above, for detecting subtle density variations it might be more suitable to use a dependence measure based on integrated functions of densities, described in the next subsections.

Integrated Functions of Density Differences

For the bivariate case, Rosenblatt [89] and Rosenblatt and Wahlen [90] considered a class of measures of dependence based on integrated squared differences of densities

$$d(f,g) = \int_{\mathbb{R}^2} w(\boldsymbol{x})(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \,,$$

for some positive weight function $w(\boldsymbol{x})$. The integral can be estimated nonparametrically by plugging in kernel density estimators for the unknown densities, and performing the integration, either numerically or, if possible, analytically. Alternatively one may estimate the integral by taking sample averages of estimated densities. For instance, if $w = 1$ one may estimate $\int_{\mathbb{R}^2} f^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^2} f(\boldsymbol{x})\mathrm{d}F(\boldsymbol{x}) = E[f(\boldsymbol{X}_t)]$ as $\tilde{n}^{-1} \sum_t \widehat{f}(\boldsymbol{X}_t)$, where $\widehat{f}(\boldsymbol{x})$ represents a consistent kernel density estimate of $f(\boldsymbol{x})$.

Chan and Tran [24] proposed a test for serial independence based on the integrated absolute difference

$$\tilde{d}(f,g) = \int_{\mathbb{R}^2} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \,, \qquad p > 0 \,,$$

for which they developed a histogram estimator [94]. They obtained critical values of the test statistic using a bootstrap method.

Skaug and Tjøstheim [96] explored tests for serial independence based on several dependence measures which are weighted integrals of $f(\boldsymbol{x}) - g(\boldsymbol{x})$ in the bivariate case ($m = 2$) including the above two measures, which they refer to as I_3 and I_2, respectively. In addition they consider the Kullback–Leibler divergence (I_1 in their notation, discussed in Subsect. "Information Theoretic Divergence Measures") and

$$I_4 = \int_{\mathbb{R}^2} (f(\boldsymbol{x}) - g(\boldsymbol{x})) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,.$$

The latter measure is not a true divergence between f and g, but if f is a bivariate normal pdf, $I_4 \geq 0$ with equality if and only if $f = g$. Skaug and Tjøstheim [96] performed a simulation study in which the corresponding estimators $\widehat{I}_i$ were compared, and $\widehat{I}_4$ was found to perform well relative to the other statistics. They subsequently investigated some of the asymptotic properties of this estimator, establishing, among other results, its asymptotic normality. Despite these encouraging simulation results one should beware that there are theoretical cases with dependence where I_4 is zero, meaning that there are also processes with dependence against which the test has little or no power.

Information Theoretic Divergence Measures

By using test statistics based on true divergences, tests can be obtained that are consistent against all deviations of f from the product measure g. Although this does not guarantee high finite sample power for specific alternatives, it obviously is a desirable property if the nature of the alternative is unknown.

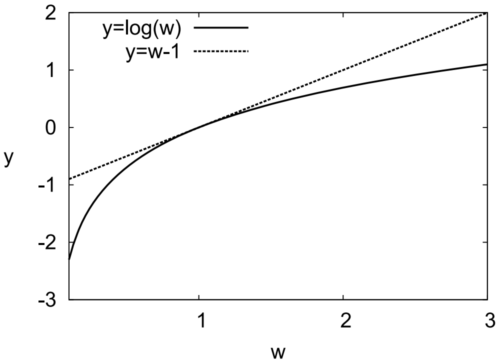
Joe [68] described several information theoretic divergence measures, including the Kullback–Leibler divergence between two densities f and g, defined as

$$I(f, g) = \int_{\mathbb{R}^m} f(\boldsymbol{x}) \log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \,.$$

In the case where f is a bivariate density, of X_{t-k} and X_t, say, and g is the product of marginal densities, $I(f, g)$ is also known as the mutual information between X_{t-k} and X_t.

Robinson [88] took the Kullback–Leibler divergence as a starting point for testing the equivalence of f and g. The Kullback–Leibler divergence is invariant under transformations of marginal distributions, and satisfies $I(f, g) \geq 0$ with equality if and only if $f = g$. To see why, consider the random variable $W = g(\boldsymbol{X})/f(\boldsymbol{X})$. By construction $E[W] = 1$, and because $\log(W)$ is a concave function of W it follows from Jensen's inequality that $E[-\log W] \leq 0$, with equality if and only if $g(\boldsymbol{X}) = f(\boldsymbol{X})$ with probability one. The reason is that $\log x \leq 1 - x$ for positive x, as illustrated in Fig 2. Application of this inequality to W shows that $E[\log(W)] \leq \log E[W] = 0$ with equality if and only if $W = 1$ with probability 1.

The fact that the Kullback–Leibler divergence is positive for any difference between the true pdf f and the hypothetical pdf g makes it a suitable quantity for testing $f = g$ against unspecified alternatives. A consistent estimator for $I(f, g)$ may serve to construct a test that is consistent (i. e. asymptotically rejects with probability one) against any fixed alternative. Robinson [88] proceeded by constructing such an estimator for $I(f, g)$ using plug-in density estimates of the unknown bivariate densities $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$. For instance, one may use the Nadaraya–Watson density estimator



Nonparametric Tests for Independence, Figure 2
Illustration of Jensen's inequality. Since the function $y = \log w$ is concave, it is bounded from above by the tangent line at $w = 1$, given by $y = w - 1$. It follows that if $E[W] = 1$ and $Y = \log W$, $E[Y] = E[\log W] \leq E[W - 1] = E[W] - 1 = 0$ with equality if and only if $W = 1$ with probability 1

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{\tilde{n} h^m} \sum_t K((\boldsymbol{x} - \boldsymbol{X}_t)/h) \,,$$

where $K(\boldsymbol{x})$ is a probability kernel, such as the pdf of a multivariate normal random variable with independent elements, $(2\pi)^{-m/2} \exp(-\boldsymbol{x}'\boldsymbol{x}/2)$, h is a smoothing parameter, and $\tilde{n}$ the number of delay vectors $\boldsymbol{X}_t$ occurring in the summation. The resulting estimator of the Kullback–Leibler divergence is

$$\widehat{I}(\widehat{f}, \widehat{g}) = \frac{1}{|S|} \sum_{t \in S} \log\left(\frac{\widehat{f}_{X_{t-k}, X_t}(X_{t-k}, X_t)}{\widehat{f}_X(X_{t-k}) \widehat{f}_X(X_t)}\right) \,, \qquad (1)$$

where S is a subset of $k + 1, \ldots, n$, introduced to allow for "trimming out" some terms of the summation if desired, for instance terms in the summation for which one or more of the local density estimates are negative or zero, as may happen depending on the type of density estimators used. The number of elements of S is denoted by $|S|$. Robinson [88] showed that although the test statistic is a consistent estimator of the Kullback–Leibler divergence, no scaled version of it has a standard normal limit distribution, preventing the development of asymptotic distribution theory in a standard fashion. To overcome this problem, instead of deriving the asymptotic distribution of $\widehat{I}(\widehat{f}, \widehat{g})$, Robinson showed asymptotic normality of a modified test statistic, obtained by attaching weights to each of the terms in the sum in (1).

Hong and White [65] argued that this modification leads to the loss of asymptotic local (i. e. close to the null) power, and developed asymptotic distribution theory for the estimator $\widehat{I}(\widehat{f}, \widehat{g})$ directly. After adjusting for the asymptotic mean they found that an appropriately scaled version of the test statistic actually does have an asymptotically standard normal distribution under the null hypothesis.

Alternatively one may obtain critical values by calculating the test statistic for a large number of simulated replications of an i.i.d. process, as done by Granger and Lin [50]. Note, however, that the critical values thus obtained will depend on the marginal distribution assumed for the process. This approach was followed by Dionísio et al. [35], who used the mutual information between X_{t-k} and X_t for a range of k-values to test for serial independence in stock index returns. Critical values of the test statistic were determined by constructing a reference distribution of the test statistic under the null hypothesis by simulation, repeatedly calculating the value of the test statistic for a large number of independently generated i.i.d. normal time series. The results suggest the presence of residual dependence at several lags for log-returns on stock indices that were filtered to account for ARMA and GARCH structure.

A closely related information theoretic approach has been described by Granger et al. [51] and Racine and Maasoumi [86], who start by considering the class of divergences based on the asymmetric q-class entropy divergence measure defined as

$$I_q(f, g) = \frac{1}{1-q}\left[1 - \int_{\mathbb{R}^m} \left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right)^q g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] .$$

This is a generalization of the Kullback–Leibler divergence, which is recovered in the limit as $q \to 1$. The authors subsequently focused on the symmetric $q = \frac{1}{2}$ case,

$$\begin{aligned} I_{\frac{1}{2}}(f, g) &= 2 - 2\int_{\mathbb{R}^m} \sqrt{g(\boldsymbol{x})}\sqrt{f(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^m} \left(\sqrt{g(\boldsymbol{x})} - \sqrt{f(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x} , \end{aligned}$$

known as the Hellinger distance, and used this to develop tests for various hypotheses involving the equality of two densities, including serial independence.

Fernandes and Neri [44] proposed using an estimator of the Tsallis entropy [105] to test for serial independence in a time series setting. As it turns out, the Tsallis entropy is identical to $I_q(f, g)$. In numerical simulation studies Fernandes and Neri [44] found that, depending on the time series processes under consideration and on the value of q, these tests can have higher power than the entropy-based test of Hong and White [65]. In comparison with the BDS test of Brock et al. [16] (see Sect. "Correlation Integrals") they found that the entropy-based tests perform worse in most cases, although the latter have more power for specific processes, including fractional AR(1) and threshold AR(1) processes.

Aparicio and Escribano [1] developed further tests based on information theoretic dependence measures. Their framework allows testing for short memory against long memory, as well as for the absence of cointegration against linear or nonlinear cointegration. In empirical applications they found that although the rates of the Peseta against the Yen and the US dollar do not appear to be linearly cointegrated, there is evidence supporting a nonlinear cointegrating relation between the two rates.

Characteristic Functions

Csörgő [26] noted that instead of investigating empirical distribution functions for testing independence, as Hoeffding [62] and Blum, Kiefer and Rosenblatt [10] did, a parallel approach can be based on empirical characteristic functions. Several tests for independence have been developed on the basis of this principle. I will be concerned here only with serial independence tests. As with the empirical distribution function one might consider various measures of deviations from independence, e. g. based on a maximum difference or on weighted integrals.

The test of Pinkse [84] is based on the observation that the random variables X_1 and X_2 are independent if and only if their joint characteristic function factorizes. He proposed to test the relation

$$\Psi(u, v) = E\mathrm{e}^{i(uX_1+vX_2)} - E\mathrm{e}^{iuX_1}E\mathrm{e}^{ivX_2} = 0$$

through a quantity of the form $\theta = \iint g(u)g(v)|\Psi(u, v)|^2 \mathrm{d}u\mathrm{d}v$, where $g(\cdot)$ is a positive function. In fact Pinkse introduced an estimator of a related but different functional, as detailed in Sect. "Quadratic Forms" where it is also explained why the test statistic can be estimated directly using U-statistics [93], without the need to actually perform the transformation to characteristic functions.

Hong [63,64] proposed a test for independence based on the difference between the joint characteristic function of X_{t-j} and X_t and the product of their marginal characteristic functions. The main idea is to weigh the discrepancy between F and G across all lags by considering the Fourier transforms

$$h(\omega, \boldsymbol{x}) := (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j(\boldsymbol{x}) \exp(-ij\omega)$$

of $\gamma_j(\boldsymbol{x}) = F_j(\boldsymbol{x}) - G(\boldsymbol{x})$ where F_j denotes the joint CDF of X_{t-j} and X_t. An application [63] to a series of weekly Deutschmark US dollar exchange rates from 1976 until 1995 showed that although the log returns are serially uncorrelated, there is evidence of nonlinear dependence of the conditional mean return given past returns.

Correlation Integrals

Correlation integrals have been used extensively in the chaos literature, where they were introduced to characterize deterministic dynamics reconstructed from time series. The interested reader is referred to Takens [99] for details of the reconstruction theorem, and to Grassberger et al. [52,53] and the book by Tong [104] for a snapshot of the early developments around correlation integrals. Correlation integrals turn out to be very suitable also in stochastic contexts. They are well adapted to testing for serial independence against unspecified alternatives, as shown below. Moreover, since they are U-statistics asymptotic theory is readily available for them [30,31].

Brock et al. [15,16] based their test for serial independence on the correlation integral of X_t^m, defined as

$$C_m(\varepsilon) = P[|\boldsymbol{Z}_1 - \boldsymbol{Z}_2| \leq \varepsilon]$$
$$\text{with } Z_i \sim \boldsymbol{X}_t^m, \text{ independent for } i = 1, 2\,,$$

where $|\cdot|$ denotes the supremum norm defined by $|\boldsymbol{x}| = \sup_{i=1,\ldots,m} |x_i|$. Under the null hypothesis of serial independence the correlation integral factorizes:

$$C_m(\varepsilon) = (C_1(\varepsilon))^m \quad . \tag{2}$$

This can be seen by expressing $C_m(\varepsilon)$ as a double integral

$$\begin{aligned} C_m(\varepsilon) &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} I_{[0,\varepsilon]}(|\boldsymbol{x}-\boldsymbol{y}|)\,\mu_m(\mathrm{d}\boldsymbol{x})\,\mu_m(\mathrm{d}\boldsymbol{y}) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_1-y_1|)\,\mu_1(\mathrm{d}x_1)\,\mu_1(\mathrm{d}y_1)\times\cdots \\ &\quad \times \int_{\mathbb{R}}\int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_1-y_1|)\,\mu_1(\mathrm{d}x_m)\,\mu_1(\mathrm{d}y_m) \\ &= (C_1(\varepsilon))^m \quad . \end{aligned}$$

In the first step the independence of $|X_i - Y_i|$ and $|X_j - Y_j|$ $(1 \leq i \neq j \leq m)$ was used, for two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ drawn independently from the distribution of $\boldsymbol{X}_t^m$ under the null hypothesis (in that case all elements $X_1, \ldots, X_m, Y_1, \ldots, Y_m$ are independent and identically distributed). Note that strictly speaking $C_m(\varepsilon) - (C_1(\varepsilon))^m$ is not a divergence. Although it will typically be nonzero for most alternatives with serial dependence, it is possible to construct examples where $C_m(\varepsilon) - (C_1(\varepsilon))^m$ is zero even under serial dependence. Formally this means that testing if $C_m(\varepsilon) - (C_1(\varepsilon))^m$ is zero, amounts to testing an implication of the null hypothesis of serial independence.

For a given kernel function $K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ that is symmetric in its arguments, the U-statistic based on a (possibly dependent) sample $\{\boldsymbol{X}_t^m\}_{t=1}^{\tilde{n}}$, consists of the sample average of the kernel function with all elements different:

$$\frac{(\tilde{n}-k)!}{\tilde{n}!} \underset{i_j \text{ all different}}{\sum_{i_1} \cdots \sum_{i_k}} K(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_k})\,.$$

The corresponding V-statistic is the sample average if the elements are allowed to be identical:

$$\frac{1}{\tilde{n}^k} \sum_{i_1} \cdots \sum_{i_k} K(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_k})\,.$$

The BDS test is based on the scaled difference between the U-statistic estimators of the left- and right-hand sides of (2):

$$W_n = \sqrt{n}\frac{C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m}{\sigma_{m,n}}\,,$$

where the U-statistic

$$C_{m,n}(\varepsilon) = \frac{2}{(n-m+1)(n-m)} \cdot \sum_{i=2}^{n-m+1} \sum_{j=1}^{i} I_{[0,\varepsilon]}(|\boldsymbol{X}_i^m - \boldsymbol{X}_j^m|)\,, \tag{3}$$

is known as the sample correlation integral at embedding dimension m and $\sigma_{m,n}^2$ is a consistent estimator of the asymptotic variance of the scaled difference. The asymptotic distribution of the test statistic can be derived using the results for U-statistics for weakly dependent processes, described by Denker and Keller [30,31]. Under the null hypothesis of serial independence,

$$W_n \xrightarrow{d} N(0,1)\,.$$

In fact the asymptotic distribution of $C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m$ is obtained from that of $(C_{m,n}(\varepsilon), C_{1,n}(\varepsilon))$. Since this is a pair of U-statistics, it follows from the results of Denker and Keller [30] that it is asymptotically bivariate normally distributed for strongly mixing stationary processes [14]. After deriving the asymptotic means and covariance matrix one can apply the functional delta method to obtain the asymptotic normal distribution of $C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m$.

To apply the BDS test the user should specify a value for the bandwidth parameter ε. In numerical studies as well as applied studies, ε-values are typically taken in the range 0.5–1.5 times the sample standard deviation of the observed time series. Note that the null hypothesis tested is independence among all elements of $\boldsymbol{X}_t^m$ rather than pairwise independence of X_{t-m+1} and X_t. Because this results in a relative error of the estimated correlation integral that increases rapidly with m, for applications with moderate sample sizes ($n \approx 1000$, say) small values of m are recommendable (e. g. $m = 2$ or $m = 3$).

Brock et al. [16] derived a 'nuisance parameter theorem' for the BDS test, showing that the limiting distribution of the test statistic is asymptotically free of estimation uncertainty of an unknown parameter θ (e. g. a vector of AR(p) model parameters) provided that a root-n consistent estimator is available for θ. The nuisance parameter theorem, which covers the parameters of AR models, but not, for instance of ARCH models, states that the asymptotic distribution of the test statistic for residuals is the same as that for the true innovations. This justifies the use of residuals in place of true innovations asymptotically, which is convenient since it allows using the BDS test on residuals as a model specification test, provided that the estimated parameters are root-n consistent.

De Lima [78] formulated five conditions under which the BDS test is asymptotically nuisance parameter free (i. e. can be used as a model specification test). These involve, among others, mixing conditions and conditions ensuring the consistency of parameter estimates. Interestingly, the test is not asymptotically nuisance parameter free for GARCH residuals, but it is when applied to logarithms of the squared residuals. Caporale et al. [21] have performed a simulation study to evaluate the behavior of the test statistic under violations of these conditions, and found the BDS test to be very robust.

Note that filtering time series data by replacing them with the (standardized) residuals of a time series model typically has the effect of whitening the data, which makes the detection of dependence more difficult. Brooks and Heravi [18] found that upon filtering data through a completely misspecified GARCH model, the frequency of rejection of the i.i.d. null hypothesis can fall dramatically. Therefore, a failure to reject the null hypothesis on the basis of GARCH residuals does not imply that a GARCH model is consistent with the data.

Wolff [106] observed that the unnormalized correlation integral, i. e. the double sum in (3) without the normalizing factor, converges to a Poisson law under some moderate assumptions regarding the marginal distribution. This motivates a nonparametric test procedure based on the correlation integral, which Wolff found to have reduced size distortion compared to the usual BDS test.

Instead of the sample correlation integral, Kočenda and Briatka [74] suggest using an estimator of the slope

$$D_m(\varepsilon) = \frac{\mathrm{d}\ln C_m(\varepsilon)}{\mathrm{d}\ln \varepsilon} \,,$$

also known as the course-grained correlation dimension at embedding dimension m and distance ε, for testing the null hypothesis of serial independence. The intuition is that the theoretically $C_m(\varepsilon) \sim \varepsilon^m$ for small ε under the i.i.d. null, while $C_m(\varepsilon) \sim \varepsilon^\alpha$ for some $\alpha < m$, provided m is sufficiently large, in the case of a low-dimensional attractor with correlation dimension α. The coarse-grained correlation dimension is a measure for complexity, and deviations from the null other than chaos typically also reduce the coarse-grained correlation dimension. This makes the coarse-grained correlation dimension a promising quantity for testing the i.i.d. null hypothesis. Rather than using the slope for a single bandwidth ε, Kočenda and Briatka [74] proposed to use an estimator of the average slope across a range of ε-values, consisting of the least squares estimator of the slope parameter β_m in the regression

$$\ln(C_{m,n}(\varepsilon_i)) = \alpha_m + \beta_m \ln(\varepsilon_i) + u_i \,, \quad i = 1, \ldots, b \,,$$

where α_m is an intercept, u_i represents an error term, and b is the number of bandwidths ε_i taken into consideration. They then determined the optimal range of ε-values by simulation. The test based on the resulting least squares estimator $\widehat{\beta}_m$ for β_m was found to have high power compared to some other tests for serial independence, and to behave well when used as a specification test.

Although it is clear that the correlation integrals from more than one bandwidth value ε_i contain more information than that from a single bandwidth, it is not clear why it would be a good idea to base a test on the estimator $\widehat{\beta}_m$. Since the correlation integral is an empirical CDF (of inter-point distances) the error terms u_i will be correlated, which typically leads to a loss of efficiency. In Subsect. "Multiple Bandwidth Permutation Tests" I discuss an alternative way to combine information from different bandwidths into a single test statistic, inspired by the rate-optimal adaptive tests of Horowitz and Spokoiny [67].

Johnson and McLelland [70,71] proposed a variation on the BDS test for testing the independence of a variable and a vector based on correlation integrals. The main idea is to test for remaining dependence between residuals and regressors, in addition to mere dependence among residuals. This might be an advisable approach in many cases, because even though theoretically a model misspecification should lead to residuals with serial dependence, it is often very hard to detect this dependence with tests on the residuals only, due to the whitening effect of the filtering.

Quadratic Forms

Quadratic forms are convenient for defining squared distances between probability distributions, which provide tests that are consistent against any type of dependence (hence including, for instance, ARCH and GARCH structure). A comparative advantage relative to the information theoretical divergences discussed in Subsect. "Information Theoretic Divergence Measures" is that they can, like correlation integrals, be estimated straightforwardly by U- and V-statistics.

The starting point for the construction of a quadratic form is a bilinear form, which may be interpreted as an inner product on the space of measures on $\mathbf{R}^m$. The quadratic forms discussed here were first applied in the context of testing for symmetries of multivariate distributions [34], and later extended to a time series context [32].

Consider, for a kernel function $K(\cdot,\cdot)$ on $\mathbb{R}^m \times \mathbb{R}^m$ the form

$$(\mu,\nu) = \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} K(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\mu(\boldsymbol{x})\mathrm{d}\nu(\boldsymbol{y})$$

for measures μ and ν. Note that this form is bilinear (linear in μ as well as ν). If this form happens to satisfy $(\mu,\mu) \geq 0$ for any (possibly signed) measure μ with $(\mu,\mu) = 0$ if and only if $\mu(A) = 0$ for all Borel subsets A of $\mathbb{R}^m$, then K is called positive definite. In the terminology introduced above, this means that $(\mu-\nu,\mu-\nu)$ is a divergence between the measures μ and ν. Note that a positive definite form defines an inner product on the space of measures on $\mathbb{R}^m$ with the usual properties:

(i) $(\mu,\nu) = (\nu,\mu)$.
(ii) $(a\mu + b\nu,\eta) = a(\mu,\eta) + b(\nu,\eta)$ for scalars a, b.
(iii) $(\mu,\mu) \geq 0$ with equality iff $\mu(A) = 0$ for any Borel subset $A \in \mathcal{A}$.

The inner product can therefore be used to define a norm of $\mu-\nu$ as $\|\mu-\nu\| = \sqrt{(\mu-\nu,\mu-\nu)}$, which satisfies all the usual properties of a distance, such as Schwarz's inequality, the triangle inequality and the parallelogram law (See e. g. Debnath and Mikusiński [27]).

In short: any positive definite kernel K defines an inner product on the space of measures on $\mathbb{R}^m$, which in turn defines a squared distance between μ and ν, given by

$$\theta = \|\mu-\nu\|^2 = (\mu-\nu,\mu-\nu)\,.$$

(For simplicity the dependence of the squared distance on the kernel function K has been suppressed in the notation.)

To pinpoint some classes of kernel functions that are suitable for our purposes (i. e. that are positive definite) let us assume that the kernel function K depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only through the difference $\boldsymbol{x}-\boldsymbol{y}$, and that the kernel function factorizes, i. e.

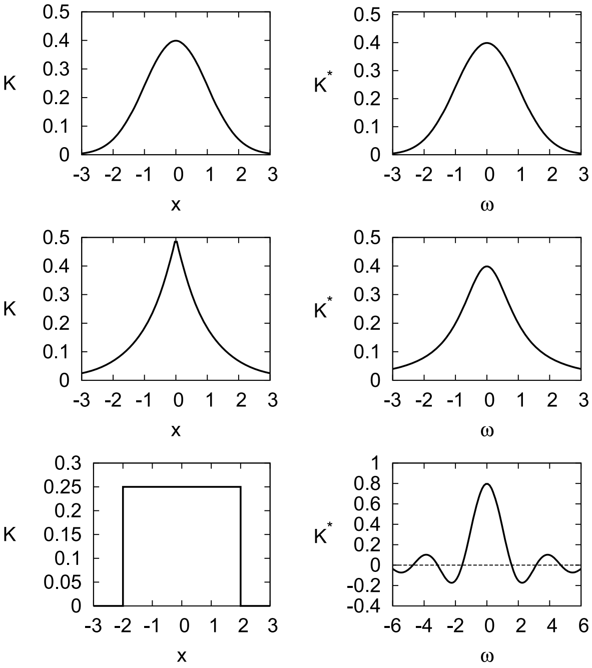
$$K(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{m} \kappa(x_i - y_i)\,.$$

In that case the Fourier transform $\tilde{K}$ of the kernel function also factorizes, into $\tilde{K}(\boldsymbol{u}) = \prod_{i=1}^m \tilde{\kappa}(u_i)$, where $\tilde{\kappa}(u) = \int \kappa(t)\mathrm{e}^{-iut}\mathrm{d}t$, the Fourier transform of κ. The squarel distance $\theta = \|\mu-\nu\|^2$ can then be expressed directly in terms of characteristic functions $\tilde{\mu}$ and $\tilde{\nu}$ of μ and ν respectively:

$$\theta = \frac{1}{2\pi}\int \tilde{K}(\boldsymbol{u})|\ \tilde{\mu}(\boldsymbol{u}) - \tilde{\nu}(\boldsymbol{u})|^2\mathrm{d}\boldsymbol{u}\,.$$

This follows from applying Parseval's theorem to $\theta = \iint K(\boldsymbol{x},\boldsymbol{y})(\mu-\nu)(\mathrm{d}\boldsymbol{x})(\mu-\nu)(\mathrm{d}\boldsymbol{y})$. It follows that if the kernel function is bounded and has a Fourier transform which does not vanish on any interval, its associated bilinear form is positive definite.

To illustrate this, Fig. 3 shows three kernel functions and their Fourier transforms. The Gaussian kernel (top panels) has a Gaussian as its Fourier transform, which is everywhere positive. Therefore, the Gaussian product



Nonparametric Tests for Independence, Figure 3
Kernel functions (*left*) and their Fourier transforms (*right*) for the Gaussian kernel (*top*), double exponential kernel (*middle*) and the naive kernel (*bottom*)

kernel is positive definite and defines a quadratic form suitable for detecting any differences between a pair of distributions on $\mathbb{R}^m$. A similar conclusion holds for the double exponential kernel $\exp(-|x|/a)$ (middle panels). The 'naive' kernel function $I_{[-a,a]}(x)$ (bottom panels) has a Fourier transform which is negative for certain ranges of the frequency ω, and hence is not a positive definite kernel function.

Given the kernel function K (e. g. a multivariate Gaussian product kernel) the estimation of the associated quadratic form $(\mu - \nu, \mu - \nu) = (\mu, \mu) - 2(\mu, \nu) + (\nu, \nu)$ is straightforward. Empirical versions of (μ, μ), (μ, ν) and (ν, ν) can be obtained easily as sample averages. For instance, if $\mathbf{X}_t^m$ is a sample from μ, the sample version of $(\mu, \mu) = \iint K(\boldsymbol{s}_1, \boldsymbol{s}_2)\mathrm{d}\mu(\boldsymbol{s}_1)\mathrm{d}\nu(\boldsymbol{s}_2)$ is the V-statistic

$$\widehat{(\mu, \mu)} = \frac{1}{\tilde{n}^2} \sum_i \sum_j K(\mathbf{X}_i^m, \mathbf{X}_j^m)\,.$$

As before, $\tilde{n} = n - m + 1$ denotes the number of m-vectors available. It follows from the results of Denker and Keller [30,31] for U- and V-statistics of dependent processes that the estimator is consistent under strong mixing conditions and asymptotically normally distributed with a variance that can be estimated consistently from the data. Note that the estimator of (μ, μ) is in fact a sample correlation integral, but with the kernel K instead of the usual naive kernel.

As shown in [32], similar consistent estimators for the other terms can be constructed easily:

$$\begin{aligned}\widehat{(\mu, \nu)} &= \frac{1}{\tilde{n}} \sum_{t=1}^{\tilde{n}} \prod_{k=0}^{m-1} \widehat{C}(X_{t+k})\,,\\ \widehat{(\nu, \nu)} &= \frac{1}{\tilde{n}^m} \prod_{k=0}^{m-1} \left(\sum_{t=1}^{n} \widehat{C}(X_{t+k}) \right)\,,\end{aligned}$$

where $\widehat{C}(x) = \frac{1}{n}\sum_{i=1}^{n} \kappa(x - X_i)$ is the one-dimensional correlation integral associated with the marginal distribution. For some results concerning size and power and comparisons of those with the BDS test and the test of Granger, Maasoumi and Racine [51] see section "Multiple Bandwidth Permutation Tests".

In fact the divergence measure $\theta = \iint g(u)g(v)$ $|\Psi(u, v)|^2 \mathrm{d}u\mathrm{d}v$, on which Pinkse [84] based his test for serial independence (see Sect. "Characteristic Functions") is also a quadratic form (for bivariate random variables). Instead of using a U-statistics estimator of θ, Pinkse used an estimator of a related quantity ϑ, which in terms of the associated inner product can be expressed as:

$$\vartheta = \left\{ [(\mu, \mu) - (\mu, \nu)]^2 + [(\nu, \nu) - (\mu, \nu)]^2 \right\} / 2\,.$$

It can be verified that also $\vartheta \geq 0$ with equality if and only if $\vartheta = 0$. Indeed, evidently $\vartheta = 0$ if $\mu = \nu$, while under any alternative one cannot have both $(\mu, \nu) = (\mu, \mu)$ and $(\mu, \nu) = (\nu, \nu)$, since in that case $(\mu - \nu, \mu - \nu) = (\mu, \mu) - 2(\mu, \nu) + (\nu, \nu) > 0$.

Tests Based on Other Measures of Dependence

Partial Sums of Data

Ashley and Patterson [2] proposed a test for independence in stock returns based on the cumulative sum $Z_t = \sum_{j=1}^{t} X_j$ where X_j represents the residuals obtained after estimating an AR(p) model on returns. The idea is that if the model is appropriate, the residuals are expected to be close to i.i.d. and Z_t corresponds to the deviation of a Brownian motion on the line after t time steps. The authors proposed to test this property using the statistic $Z^{\max} = \max\{|Z_1|, \ldots, |Z_T|\}$, assessing the statistical significance using a bootstrap method.

It was later pointed out by Corrado and Schatzberg [25] that since $\{X_t\}$ has a zero sample mean, $\{Z_t\}$ is 'tied to zero' at the endpoints ($Z_t = Z_0 = 0$), and hence the reference paths used in the bootstrap should have been constructed to satisfy the same constraints. This can, for instance, be achieved by mean adjusting the bootstrap sample, or alternatively by employing a permutation method (resampling without replacement). Moreover, Corrado and Schatzberg [25] showed that after rescaling via

$$W_t = Z_t / (\sqrt{T}\widehat{\sigma}_T)$$

where $\widehat{\sigma}_T$ is the sample standard deviation of X_j, the sample path of W_t for large T forms a Brownian bridge under the null hypothesis, which implies that the maximum absolute value has the same null distribution as the Kolmogorov–Smirnov (KS) test statistic.

Scaled versions of partial sums were also considered by Kulperger and Lockhart [75]. They focus on examining the conditional mean of $Y_j := X_{j+1}$ given X_j, by studying the dependence among successive Y-values when ordered according to the ranks of the corresponding X-values. Put simply, this replaces the ordering of pairs (X_t, Y_t) in such a way that X-values are ordered increasingly, enabling the partial sums to grasp the common dependence between Y-values on X-values, rather than on time. Motivated by this, the authors propose to study the sample path of the partial sums

$$S_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{i} (Y_{(j)} - \bar{Y})\,,$$

where $Y_{(j)}$ denotes $X_{(j)+1}$ (the successor of the observation among whose rank among the original observations is j), and $\bar{Y}$ is the sample mean of $\{Y_j\}$. The authors then propose and compare various statistics to test if the realized process $\{S_i\}$ is a realization of a Brownian bridge, as predicted under the null hypothesis. Straightforward extensions can be obtained by taking $Y_{(j)} = \Phi(X_{(j)+k})$ for some fixed lag k.

The Spectral Density

Besides being able to test the strict random walk hypothesis (i.i.d. increments) for a financial time series such as a log-price, it is also of interest to be able to test the weaker hypothesis that increments have a constant conditional mean. A test based on the spectral density for this so-called martingale hypothesis was developed by Durlauf [41].

The Bispectrum

As already noted by Robinson [88], one can test for serial independence against nonlinear dependence with a test for linearity rather than independence. Here and in the next subsection I briefly discuss a few examples of linearity tests.

Extending results of Subba-Rao [98], Hinich [58] used the bispectrum to detect 'interactions between Fourier components'. The motivation behind the approach is that the bicorrelation, defined as

$$c(k,\ell) = E[X_t X_{t+k} X_{t+\ell}] \, ,$$

should be zero for a stationary linear Gaussian random process $\{X_t\}$ with mean zero.

As an illustration of the structure that the bicovariance of a nonlinear time series may exhibit, Fig. 4 shows the bicovariance for the time series $\{X_t\}$ generated by the bilinear process

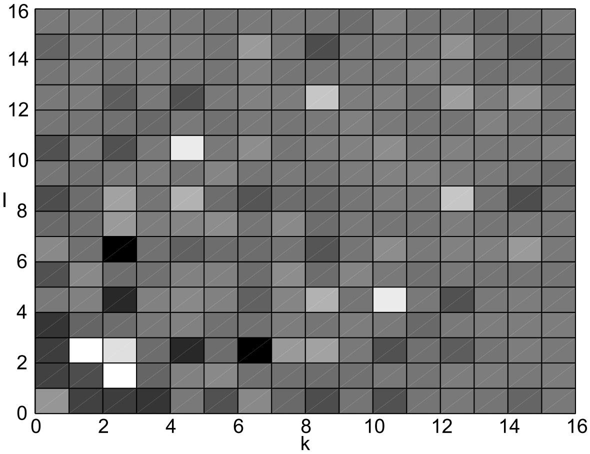
$$X_t = 0.9\varepsilon_{t-1}X_{t-2} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. For the simulated data I initialized the state variables at $X_{-1} = X_0 = 0$ and discarded the first 1000 iterations.

Examining the behavior of the bicorrelation $c(k,\ell)$ for many values of k and ℓ simultaneously can be achieved in various ways, for instance by examining the bispectrum

$$B(\omega_1,\omega_2) = \sum_{k=-\infty}^{\infty}\sum_{\ell=-\infty}^{\infty} c(k,\ell)\exp\left[-i(\omega_1 k + \omega_2 \ell)\right] .$$

Hinich [58] introduced two functionals of the bispectrum that are suitable for testing Gaussianity and linearity, respectively. For applications to both model generated data and real data see, among others, [3,4,17,57]. Overall, these applications indicate that nonlinearity and non-Gaussianity play an important role in economic time series.



Nonparametric Tests for Independence, Figure 4
Bicovariance $c(k,\ell)$ for a bilinear process. *Lighter regions* correspond with larger values of the bicovariance. Series length $n = 4000$

More recently Hinich [59] proposed a related test for serial independence of the innovations acting on a (unspecified) linear filter. This test is also based on the bispectrum, but the test statistic can be evaluated in the time domain since it is a function of the sample bicovariance function of the residuals. Lim et al. [77] applied the test of Hinich [59] to returns of Asian market indices. If the returns would follow a GARCH process with symmetric innovations the sequence of signs thus obtained should be i.i.d. Since the results indicate nonlinear structure in the signs of Asian stock returns, the conclusion is that GARCH models with symmetric innovations are inappropriate for the returns.

Note that although this example shows how the bispectrum can be used to detect evidence against the null of (G)ARCH, the bispectrum cannot be used to test for serial independence against (G)ARCH alternatives. The reason is that the bicovariance is not able to pick up dependence in time series processes in which the conditional mean of X_t given past observations is zero, such as ARCH and GARCH processes.

If the probabilistic structure of a stationary time series process is preserved under time-reversal (i. e. when reading the series backward in time), the time series process is called time reversible. Clearly, time reversibility should hold under serial independence. Ramsey and Rothman [87,91] developed a test for time reversibility based on a sample version of the difference of bicovariances $c(k,k) - c(0,k) = E[X_t^2 X_{t-k}] - E[X_t X_{t-k}^2]$. This test is consistent against some forms of serial dependence, but not against all. For instance, it is not consistent against ARCH and GARCH alternatives, since these have zero bicovariance at any lag k.

Terdik and Máth [100] use the information contained in the bispectrum to test whether the best predictor is linear against quadratic alternatives. The null hypothesis being tested is linearity in mean, as opposed to a linear Gaussian random processes.

Brooks and Hinich [19,20] generalized the bispectrum approach to multivariate time series settings, in order to study nonlinear lead-lag relationships. These authors indeed found evidence for nonlinear dependence structure between various exchange rates. As noted by the authors these findings have important implications for investors who try to diversify their portfolios internationally.

Nonlinearity in the Conditional Mean

Hjellvik and Tjøstheim [60] developed a nonparametric test for linearity based on the difference between the best linear predictor and a kernel-based nonparametric predictor. Hjellvik et al. [61] explored a variant of this approach where local polynomial predictors were used instead of kernel-based predictors.

Bootstrap and Permutation Tests

As shown above, in many cases it is possible to use asymptotic distribution theory for test statistics in the time series context. Notable cases are those where the test statistics are U-statistics or a function thereof, as is the case for the BDS test. In some cases, however, the resulting asymptotic approximation to the finite sample null distribution may be poor. In particular this can happen when the test statistic is a degenerate or near-degenerate U-statistic under the null hypothesis. For instance, the BDS test statistic is near-degenerate under the null hypothesis if the marginal distribution is uniform.[1]

For practical purposes, near-degeneracy means that although the test statistic is asymptotically standard normal, it may be far from asymptotic normality even for large sample sizes. Whether a particular test statistic is (near) degenerate under the null is often not known, as it depends on the marginal distribution of the data generating process. To avoid such problems with the asymptotic approximation one can use critical values obtained by simulating a fully specified model that satisfies the i.i.d. null hypothesis. However, since the distribution of the test statistic depends on the marginal distribution, it is better to reflect this in the simulated data as well. This can be done by using bootstrap or Monte Carlo methods for assessing the statistical significance of the observed value of the test statistic of interest. The Monte Carlo procedure has the additional advantage that it produces an exact randomization test.

Simulation

Although I do not recommend this procedure in practice, I include it for completeness. For simplicity I describe the approach here only for the one-sided case where large values of the test statistic provide evidence against the null hypothesis. The approach for two-sided tests is similar.

Suppose we wish to obtain critical values of a test statistic, Q_n, say, then this can be obtained by simulating a large number of i.i.d. time series of length n. The idea is to simulate time series satisfying the null hypothesis using a fully specified model. For instance one can simulate a large number B of i.i.d. normal time series data of the same length as the original time series, and then calculate the value of the test statistic for each of these artificial time series. The sample distribution of the B simulated test statistics subsequently represents the null distribution, and a p-value can be obtained as $\widehat{p} = \#\{Q_n^i \geq Q_n\}/(B+1)$. This approach is suitable if one is willing to assume a certain marginal (normality in this case) under the null, or if the distribution of the test statistic is (at least asymptotically) independent of the (unknown) marginal distribution.

Dionísio [35] implemented a test for serial independence based on the mutual information between X_t and $X_{t+\tau}$. Critical values of the test statistic were obtained for individual lags τ by simulating a large number of $N(0,1)$ i.i.d. time series of the appropriate length. This should provide a good approximation to the true critical values if the data are approximately normally distributed, or if the the (asymptotic) distribution of the test statistic is independent of the (unknown) marginal distribution. If not, this may lead to size distortions if the data are skewed or otherwise deviating from normality. A bootstrap or a permutation test may be more appropriate in those cases.

Bootstrap Tests

The bootstrap approach consists of resampling from the observed data with replacement. The idea is that under the

[1] It is exactly degenerate for i.i.d. data from the uniform distribution on the circle, i. e. the interval $[0, a]$ with the endpoints identified [16]. Theiler [102] simulated variance of $S = C_{m,n}(\varepsilon) - (C_{i,n}(\varepsilon))^m$ in the degenerate case, and found that it converges to 0 at the rate n^{-2} instead of the usual rate n^{-1}.

null the best representation of the data generating process that we have is given by an i.i.d. process with the empirical distribution of the observed data. One of the motivating advantages of the bootstrap is that it yields an improved approximation to the finite sample distribution of test statistics relative to first-order asymptotic theory, provided that the statistics are asymptotically pivotal. For an overview of the use of the bootstrap in econometrics, I refer the interested reader to [66].

Although there are sophisticated bootstrap methods that are particularly designed for time series, for instance the block bootstrap [22,76,85], in the case of testing for serial independence the data are i.i.d. under the null, so under the null hypothesis we can bootstrap by simply drawing n values from the original time series independently with replacement. This procedure is often referred to as the naive bootstrap.

Hong and White [65] noted that the naive bootstrap does not produce a consistent procedure for their test statistic (essentially the estimator of the Kullback–Leibler divergence given in (1)) as it is degenerate under the null hypothesis of serial independence. They propose the use of a smoothed bootstrap procedure to overcome this. In the degenerate case also a permutation test may be used, as described next.

Permutation Tests

Under the null hypothesis of serial independence the data generating process is typically known only up to an infinite dimensional nuisance parameter (the unknown marginal distribution). This prevents one from generating time series data that have the exact same distribution as the data generating process under the null hypothesis, as Barnard [5] suggested for simple null hypotheses (i. e. null hypotheses under which the distribution of the data is fully specified). Hence the problem is that the null hypothesis of serial independence is not simple but composite, with each possible marginal distribution representing another i.i.d. process. This limitation can be overcome by considering all the possible processes under the null, conditional on an observed value of a minimal sufficient statistic for the nuisance parameter. The resulting (conditional) null distribution of the test statistic can then be shown to be free of unknown parameters. The simulated data should be drawn from the same conditional distribution as the data generating process under the null, given the sufficient statistic. This procedure can be used for constructing a randomization test procedure which is exact, i. e. the type I error rate is equal to nominal level, at least in the absence of parameter estimation uncertainty. The resulting tests are referred to as Monte Carlo tests.

Since under the null hypothesis the empirical marginal distribution is a minimal sufficient statistic for the unknown marginal distribution. The conditional distribution of the observations given their empirical marginal, assigns equal probability to each of the $n!$ possible permutations of the data. This means that every permutation of the originally observed values is equally likely. Hence an independent draw from the time series process conditional on the sufficient statistic can be obtained straightforwardly by randomly permuting the original data. The value of the test statistic for such a permuted time series is an independent draw from the conditional null distribution of the test statistic given the sufficient statistic. Although the Monte Carlo method is exact for data generated under the null hypothesis, not many investigators have studied its behavior when applied to residuals of an estimated time series model. For a general treatment of Monte Carlo testing in the presence of model parameter uncertainty, see the recent work by Dufour [37].

Multiple Bandwidth Permutation Tests

Most of the nonparametric tests described above, such as the BDS test, require a user-specified value for the bandwidth parameter. I mentioned that the BDS test is usually applied with bandwidth values in the range 0.5 to 1.5 standard deviations. Although these values appear to work reasonably well in numerical simulation studies with computer-generated data from known processes, there is no guarantee that this is an optimal range for the (usually unknown) alternative of most interest to the user (i. e. the possibly non-i.i.d. true process that generated the data at hand). In the different context of testing a parametric regression function against an unknown nonparametric alternative, Horowitz and Spokoiny [66] proposed tests with an adaptive bandwidth, that they showed to be rate-optimal. Although the present context is slightly different, and the details of their theorems most likely require some adaptation before they apply here, test statistics similar to the adaptive bandwidth statistics that they proposed can be easily implemented.

The idea is to calculate test statistics for many values of the bandwidth parameter, ε, say, and reject the null hypothesis if there is evidence against independence from one or more of the statistics calculated for the various bandwidths. To achieve this, an overall test statistic is required that will pick up evidence of dependence from any of the bandwidths. In Ref. [32] we proposed us-

Nonparametric Tests for Independence, Table 1

Observed rejection rates at nominal size 0.05 of the test of Granger, Maasoumi and Racine [51] (GMR) test and the multiple bandwidth permutation procedure for the BDS test statistic [16] and the quadratic form-based test of Diks and Panchenko [32] (DP). In the model specifications $\{u_t\}$ represents a sequence of independent standard normal random variables. Embedding dimension $m = 3$, sample size 100, except for the sign process (sample size 50) and the logistic and the Hénon maps (sample size 20). Monte Carlo parameters $B = 100$ permutations and 1000 independently realized time series from each process

Model	Specification	GMR	BDS	DP
1	$X_t = u_t$	0.05	0.05	0.05
2	$X_t = u_t + 0.8u_{t-1}^2$	0.57	0.68	0.71
3	$X_t = u_t + 0.6u_{t-1}^2 + 0.6u_{t-2}^2$	0.78	0.84	0.96
4	$X_t = u_t + 0.8u_{t-1}u_{t-2}$	0.22	0.46	0.29
5	$X_t = 0.3X_{t-1} + u_t$	0.31	0.16	0.68
6	$X_t = 0.8\|X_{t-1}\|^{0.5} + u_t$	0.25	0.11	0.53
7	$X_t = \mathrm{sign}(X_{t-1}) + u_t$	0.86	0.75	0.98
8	$X_t = 0.6\varepsilon_{t-1}X_{t-2} + \varepsilon_t$	0.26	0.50	0.39
9	$X_t = \sqrt{h_t}u_t,\ h_t = 1 + 0.4X_{t-1}^2$	0.26	0.51	0.24
10	$X_t = \sqrt{h_t}u_t,\ h_t = 0.01 + 0.80h_{t-1} + 0.15X_{t-1}^2$	0.15	0.35	0.18
11	$Y_t = (-0.5 + 0.9I_{[0,\infty)}(Y_{t-1}))Y_{t-1} + \varepsilon_t$	0.34	0.06	0.87
12	$X_t = 4X_{t-1}(1 - X_{t-1}), \quad (0 < X_t < 1)$	0.95	0.71	0.90
13	$X_t = 1 + 0.3X_{t-2} - 1.4X_{t-1}^2$	0.96	0.46	0.97
14	$X_t = Z_t + \sigma u_t, \quad Z_t = 1 + 0.3Z_{t-2} - 1.4Z_{t-1}^2$	0.41	0.22	0.83

ing the smallest p-value, $\widehat{p}(\varepsilon_i)$, across a set of d different bandwidths $\varepsilon_1 < \cdots < \varepsilon_d$ as an overall test statistic: $T = \inf_i \widehat{p}(\varepsilon_i)$. To establish if the value of T obtained is significant, a permutation test can be performed. I refer to this procedure as the multiple bandwidth permutation test.

Suppose that we wish to base the p-values $\widehat{p}(\varepsilon_i)$ on the permutation procedure described above, then this setup seems to require two nested permutation procedures; one global loop for replicating B values of T, T_i, $i = 1, \ldots, B$, for the B different permutations of the original data, and for each of those another loop to obtain a p-values $\widehat{p}(\varepsilon_i)$ of the observed (BDS) test statistic for each bandwidth. It turns out, however, that this can be achieved much more efficiently, in a single loop across B permutations of the original data, as follows.

Let $Q^1(\varepsilon_i)$, $i = 1, \ldots, b$, denote the value of the (BDS) test statistic for the original data at the ith bandwidth, ε_i, and $Q^k(\varepsilon_i)$, $k = 2, \ldots, B$, that of the kth randomly permuted time series, then a p-value can be obtained for each bandwidth as before: $\widehat{p}^1(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^1(\varepsilon_i)\}/B$. The superscript 1 denotes that these p-values are obtained for the original time series, $b = 1$. Subsequently one can obtain similar p-values for each of the permuted time series as $\widehat{p}^b(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^b(\varepsilon_i)\}/B$. Now we are in a position to calculate the global test statistic $T_b = \inf_i \widehat{p}^b(\varepsilon_i)$ for each of the B permuted time series, including the original time series (the case $b = 1$). Finally, we can establish the significance of the test statistic T_1 obtained for the original time series by comparing it with the reference values $T_2, \ldots, T_B$. Although these values need not be independent, even under the null hypothesis, they do satisfy permutation symmetry under the null hypothesis, so that each of the possible permutations of the observed values of T_b is equally likely. By the permutation symmetry of all the time series (the original and the permuted series) under the null hypothesis, and hence of the values T_b, $b = 1, \ldots, B$, the overall p-value can still be calculated as if the values T_b were independent, i. e. $\widehat{p} = \#\{T_s \leq T_1\}/B$. In other words, only the fact that all possible orderings of the values $T_1, \ldots, T_B$ are equally likely under the null hypothesis is needed, and not their independence.

So far I haven't discussed the possibility that ties may occur. In fact they occur with nonzero probability since $\widehat{p}^b(\varepsilon_i)$ is a discrete random variable for finite B. If ties are dealt with appropriately, however, then the above procedure leads to a test with a rejection rate under the null hypothesis equal to the nominal size (for details see [32]).

Table 1 shows the power of the multiple bandwidth procedure for the BDS test [16] and the test developed by Valentyn Panchenko and me [32] based on quadratic forms for various processes (referred to henceforth as the DP test). For comparison the power of the test of Granger Maasoumi and Racine [51] (GMR test) based on the Hellinger distance, discussed in Subsect. "Information Theoretic Divergence Measures", are also provided. The GMR test was performed with the R software provided by the authors, which uses a bandwidth based on cross-validation.

The processes are, in order, models of type: i.i.d. normal (1), nonlinear moving average (2–4), linear autoregressive (5), nonlinear autoregressive (6), sign autoregressive (7), bilinear (8), ARCH(1) (9), GARCH(1,1) (10), threshold autoregressive (11), logistic map (12), Hénon map (13) and the Hénon map with dynamic noise (14). The multiple bandwidth permutation test was performed with 5 bandwidth values ε_i between $\varepsilon = 0.5$ and 2.0, with a constant ratio $\varepsilon_{i+1}/\varepsilon_i$, $i = 2, \ldots, 4$ (hence the bandwidths are equally spaced on a logarithmic scale).

The table shows rejection rates for the i.i.d. process which are all close to the nominal size 0.05, hence there is no evidence for size distortion for any of the three tests. In terms of power (remaining processes) none of the tests does uniformly outperform the others, even within model classes such as the nonlinear moving average processes considered (process 2–4). This emphasizes again how hard it is to tell beforehand which test will perform best for an unknown alternative.

For applications I would have a slight preference for using a test that is consistent against any fixed alternative (such as the DP test), if only to hedge against the possibility of having no asymptotic power. However, as Table 1 shows, this does not guarantee a good finite sample performance in all cases.

Future Directions

Although permutation tests have been shown to have the advantage of providing exact tests in the ideal case of data that are truly i.i.d. under the null hypothesis, more work is required to establish the properties of these (or adapted) tests in the presence of residuals of an estimated parametric model. This requires either adaptation of the permutation procedure in that setting, or analogues of the 'nuisance parameter theorem' for the BDS test.

Another remaining challenge is the detection of dependence within observed high-variate vector-valued time series. Estimating (functionals of) probability densities in high-dimensional spaces is notoriously difficult, since the number of observations typically required grows very fast with the number of dimensions. Due to this so-called curse of dimensionality, the kernel-based methods discussed here in practice cannot be meaningfully applied to data sets with moderate sample sizes (several thousand observations) if the dimension m exceeds 5 or 6.

Additional work is also required for the development of statistical tests for time series that do not take values in the real line, but in more general manifolds. As mentioned in the introduction, an example consists of wind direction data taking values in the interval $[0, 2\pi]$ with the endpoints identified (i. e. on the circle). This touches upon the general problem of defining divergence measures between distributions on less familiar measurable spaces, and constructing and studying the statistical properties of their estimators.

Bibliography

1. Aparicio FM, Escribano A (1998) Information-theoretic analysis of serial dependence and cointegration. Stud nonlinear dyn econ 3:119–140
2. Ashley RA, Patterson DM (1986) A nonparametric, distribution-free test for serial independence in stock returns. J Financial Quant Anal 21:221–227
3. Ashley RA, Patterson DM (1989) Linear versus nonlinear macroeconomics: A statistical test. Int Econ Rev 30:165–187
4. Ashley RA, Patterson DM, Hinich MN (1986) A diagnostic check for nonlinear serial dependence in time series fitting errors. J Time Ser Anal 7:165–187
5. Barnard GA (1963) Discussion of Professor Bartlett's paper. J Royal Stat Soc Ser B 25:294
6. Bartels R (1982) The rank version of von Neumann's ratio test for randomness. J Am Stat Assoc 77:40–46
7. Benghabrit Y, Hallin M (1992) Optimal rank-based tests against 1st-order superdiagonal bilinear dependence. J Stat Plan Inference 32:45–61
8. Bera AK, Robinson PM (1989) Tests for serial dependence and other specification analysis in models of markets in equilibrium. J Bus Econ Stat 7:343–352
9. Beran J (1992) A goodness-of-fit test for time-series with long-range dependence. J Royal Stat Soc Ser B 54:749–760
10. Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free tests of independence based on sample distribution functions. Ann Math Stat 32:485–498
11. Bollerslev T (1986) Generalized autoregressive heteroskedasticity. J Econometrics 31:307–327
12. Booth GG, Martikainen T (1994) Nonlinear dependence in Finnish stock returns. Eur J Oper Res 74:273–283
13. Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 332:1509–1526
14. Bradley R (1986) Basic properties of strong mixing conditions. In: Eberlein E, Taqqu MS (eds) Dependence in Probability and Statistics. Birkäuser, Basel
15. Brock WA, Dechert WD, Scheinkman JA (1987) A test for independence based on the correlation dimension. Working paper 8702. University of Wisconsin, Madison
16. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econometric Rev 15:197–235
17. Brockett PL, Hinich MD, Patterson D (1988) Bispectral based tests for the detection of Gaussianity and linearity in time series. J Am Stat Assoc 83:657–664

18. Brooks C, Heravi SM (1999) The effect of (mis-specified) GARCH filters on the filite sample distribution of the BDS test. Comput Econ 13:147–162
19. Brooks C, Hinich MJ (1999) Cross-correlations and cross-bicorrelations in Sterling exchange rates. J Empir Finance 6:385–404
20. Brooks C, Hinich MJ (2001) Bicorrelations and cross-bicorrelations as non-linearity tests and tools for exchange rate forecasting. J Forecast 20:181–196
21. Caporale GM, Ntantamis C, Pantelidis T, Pittis N (2005) The BDS test as a test for the adequacy of a GARCH(1,1) specification: A Monte Carlo study. J Financial Econometric 3:282–309
22. Carlstein E (1984) The use of sub-series methods for estimating the variance of a general statistic from a stationary time series. Ann Stat 14:1171–1179
23. Carlstein E (1988) Degenerate U-statistics based on non-independent observations. Calcutta Stat Assoc Bull 37:55–65
24. Chan NH, Tran LT (1992) Nonparametric tests for serial independence. J Time Ser Anal 13:19–28
25. Corrado CJ, Schatzberg J (1990) A nonparametric, distribution-free test for serial independence in stock returns: A correction. J Financial Quant Anal 25:411–415
26. Csörgő S (1985) Testing for independence by the empirical characteristic function. J Multivar Anal 16:290–299
27. Debnath L, Mikusiński P (2005) Introduction to Hilbert Spaces With Applications, 3rd edn. Elsevier Academic Press, Burlington
28. Delgado M, Mora J (2000) A nonparametric test for serial independence of regression errors. Biometrika 87:228–234
29. Delgado MA (1996) Testing serial independence using the sample distribution function. J Time Ser Anal 11:271–285
30. Denker M, Keller G (1983) On U-statistics and v. Mises' statistics for weakly dependent processes. Z Wahrscheinlichkeitstheorie verwandte Geb 64:505–522
31. Denker M, Keller G (1986) Rigorous statistical procedures for data from dynamical systems. J Stat Phys 44:67–93
32. Diks C, Panchenko V (2007) Nonparametric tests for serial independence based on quadratic forms. Statistica Sin 17:81–97
33. Diks C, Panchenko V (2008) Rank-based entropy tests for serial independence. Stud Nonlinear Dyn Econom 12(1)art.2:0–19
34. Diks C, Tong H (1999) A test for symmetries of multivariate probability distributions. Biometrika 86:605–614
35. Dionísio A, Menezes R, Mendes DA (2006) Entropy-based independence test. Nonlinear Dyn 44:351–357
36. Dufour JM (1981) Rank tests for serial dependence. J Time Ser Anal 2:117–128
37. Dufour JM (2006) Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. J Econom 133:443–477
38. Durbin J, Watson GS (1950) Testing for serial correlation in least-squares regression, I. Biometrika 37:409–428
39. Durbin J, Watson GS (1951) Testing for serial correlation in least-squares regression, II. Biometrika 38:159–177
40. Durbin J, Watson GS (1971) Testing for serial correlation in least-squares regression, III. Biometrika 58:1–19
41. Durlauf S (1991) Spectral based testing of the martingale hypothesis. J Econometrics 50:355–376
42. Engle R (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
43. Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. Can J Stat 28:587–604
44. Fernandes M, Neri B (2008) Nonparametric entropy-based tests of independence between stochastic processes. Econometric Reviews; Forthcoming
45. Genest C, Quessy JF, Rémillard B (2002) Tests of serial independence based on Kendall's process. Can J Stat 30:1–21
46. Genest C, Rémillard B (2004) Tests of independence and randomness based on the empirical copula process. Test 13:335–369
47. Genest C, Verret F (2005) Locally most powerful rank tests of independence for copula models. Nonparametric Stat 17:521–539
48. Genest C, Ghoudi K, Rémillard B (2007) Rank-based extensions of the Brock, Dechert, and Scheinkman test. J Am Stat Assoc 102:1363–1376
49. Ghoudi K, Kulperger RJ, Rémillard B (2001) A nonparametric test of serial independence for time series and residuals. J Multivar Anal 79:191–218
50. Granger C, Lin JL (2001) Using the mutual information coefficient to identify lags in nonlinear models. J Time Ser Anal 15:371–384
51. Granger CW, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. J Time Ser Anal 25:649–669
52. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
53. Grassberger P, Schreiber T, Schaffrath C (1991) Nonlinear time sequence analysis. Int J Bifurc Chaos 1:521–547
54. Hallin M, Ingenbleek J-F, Puri ML (1985) Linear serial rank tests for randomness against ARMA alternatives. Ann Stat 13:1156–1181
55. Hallin M, Mélard G (1988) Rank-based tests for randomness against first-order serial dependence. J Am Stat Assoc 83:1117–1128
56. Hannan EJ (1957) Testing for serial correlation in least squares regression. Biometrika 44:57–66
57. Hinich M, Patterson D (1985) Evidence of nonlinearity in stock returns. J Bus Econ Stat 3:69–77
58. Hinich MJ (1982) Testing for Gaussianity and linearity of a stationary time series. J Time Ser Anal 3:169–176
59. Hinich MJ (1996) Testing for dependence in the input to a linear time series model. J Nonparametric Stat 8:205–221
60. Hjellvik V, Tjøstheim D (1995) Nonparametric tests of linearity for time series. Biometrika 82:351–368
61. Hjellvik V, Yao Q, Tjøstheim D (1998) Linearity testing using polynomial approximation. J Stat Plan Inference 68:295–321
62. Hoeffding W (1948) A non-parametric test of independence. Ann Math Stat 19:546–557
63. Hong Y (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. J Am Stat Assoc 94:1201–1220
64. Hong Y (2000) Generalized spectral tests for serial dependence. J Royal Stat Soc Ser B 62:557–574
65. Hong Y, White H (2005) Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Econometrica 73:837–901

66. Horowitz JL (2001) The bootstrap. In: Heckman JJ, Leamer EE (eds) Handbook of Econometrics, vol 5. Elsevier, Amsterdam, pp 3159–3228
67. Horowitz JL, Spokoiny VG (2001) An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69:599–631
68. Joe H (1989) Relative entropy measures of multivariate dependence. J Am Stat Assoc 84:157–164
69. Joe H (1990) Multivariate concordance. J Multivar Anal 35:12–30
70. Johnson D, McLelland R (1997) Nonparametric tests for the independence of regressors and disturbances as specification tests. Rev Econ Stat 79:335–340
71. Johnson D, McLelland R (1998) A general dependence test and applications. J Appl Econometrics 13:627–644
72. Kallenberg WCM, Ledwina T (1999) Data driven rank tests for independence. J Am Stat Assoc 94:285–301
73. Kendall MG (1938) A new measure of rank correlation. Biometrika 30:81–93
74. Kočenda E, Briatka Ľ (2005) Optimal range for the IID test based on integration across the correlation integral. Econometric Rev 24:265–296
75. Kulperger RJ, Lockhart RA (1998) Tests of independence in time series. J Time Ser Anal 1998:165–185
76. Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1241
77. Lim KP, Hinich MJ, Liew VKS (2005) Statistical inadequacy of GARCH models for Asian stock markets: Evidence and implications. J Emerg Mark Finance 4:263–279
78. Lima P De (1996) Nuisance parameter free properties of correlation integral based statistics. Econometric Rev 15:237–259
79. Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65:297–202
80. Lo AW (2000) Finance: A selective survey. J Am Stat Assoc 95:629–635
81. Maasoumi E (2002) Entropy and predictability of stock market returns. J Econometrics 107:291–312
82. McLeod AI, Li WK (1983) Diagnostic checking ARMA time series models using squared-residual autocorrelations. J Time Ser Anal 4:269–273
83. Von Neumann J (1941) Distribution of the ratio of the mean square successive difference to the variance. Ann Math Stat 12:367–395
84. Pinkse J (1998) A consistent nonparametric test for serial independence. J Econometrics 84:205–231
85. Politis DN, Romano JP (1994) The stationary bootstrap. J Am Stat Assoc 89:1303–1313
86. Racine J, Maasoumi E (2007) A versatile and robust metric entropy test for time-irreversibility, and other hypotheses. J Econometrics 138:547–567
87. Ramsey JB, Rothman P (1990) Time irreversibility of stationary time series: estimators and test statistics. Unpublished manuscript, Department of Economics, New York University and University of Delaware
88. Robinson PM (1991) Consistent nonparametric entropy-based testing. Rev Econ Stud 58:437–453
89. Rosenblatt M (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann Stat 3:1–14
90. Rosenblatt M, Wahlen BE (1992) A nonparametric measure of independence under a hypothesis of independent components. Stat Probab Lett 15:245–252
91. Rothman P (1992) The comparative power of the TR test against simple threshold models. J Appl Econometrics 7:S187–S195
92. Scaillet O (2005) A Kolmogorov–Smirnov type test for positive quadrant dependence. Can J Stat 33:415–427
93. Serfling RJ (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York
94. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York
95. Skaug HJ, Tjøstheim D (1993a) A nonparametric test for serial independence based on the empirical distribution function. Biometrika 80:591–602
96. Skaug HJ, Tjøstheim D (1993b) Nonparametric tests of serial independence. In: Subba Rao T (ed) Developments in Time Series Analysis: the M. B. Priestley Birthday Volume. Wiley, New York
97. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101
98. Subba Rao T, Gabr MM (1980) A test for linearity of stationary time series. J Time Ser Anal 1:145–158
99. Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young LS (eds) Dynamical Systems and Turbulence, Warwick 1980. (Lecture Notes in Mathematics), vol 898. Springer, Berlin, pp 366–381
100. Terdik G, Máth J (1998) A new test of linearity of time series based on the bispectrum. J Time Ser Anal 19:737–753
101. Theil H, Nagar AL (1961) Testing the independence of regression disturbances. J Am Stat Assoc 56:793–806
102. Theiler J (1990) Statistical precision of dimension estimators. Phys Rev A 41:3038–3051
103. Tjøstheim D (1996) Measures of dependence and tests of independence. Statistics 28:249–284
104. Tong H (1990) Non-linear Time Series: A Dynamical Systems Approach. Clarendon Press, Oxford
105. Tsallis C (1998) Generalized entropy-based criterion for consistent testing. Phys Rev E 58:1442–1445
106. Wolff RC (1994) Independence in time series: another look at the BDS test. Philos Trans Royal Soc Ser A 348:383–395

Public Policy, System Dynamics Applications to

DAVID F. ANDERSEN[1], ELIOT RICH[2], RODERICK MACDONALD[3]

[1] Rockefeller College of Public Affairs and Policy, University at Albany, Albany, USA
[2] School of Business, University at Albany, Albany, USA
[3] Initiative for System Dynamics in the Public Sector, University at Albany, Albany, USA

Article Outline

Glossary

Causal loop diagram A diagrammatic artifact that captures the causal model and feedback structure underlying a problem situation. Commonly used as a first-cut tool to identify major stakeholder concerns and interactions. These diagrams are often precursors to formal models.

Dynamic modeling Formal examination of the behavior of a system over time. Contrast with point-estimation, which attempts to predict an average outcome.

Feedback A relationship where two or more variables are linked over time so that the influence of one variable on a second will later affect the state of the first. If the influence is such as to increase the state of the first over time, the feedback is termed reinforcing. If the influence is such as to decrease the state of the first, it is termed balancing.

Formal model The representation of a system structure in mathematical form. Contrast with causal model, which represents structure without the underlying mathematics.

Mental model The representation of a problem's structure as possessed by an expert in a particular domain. Mental models are often intangible until explicated by the expert.

Public policy Any and all actions or non-actions, decisions or non-decisions taken by government, at all levels, to address problems. These actions, non-actions, decisions or non-decisions are implemented through laws, regulations and the allocation of resources.

Group model building (GMB) An approach to problem definition that asks multiple experts and major stakeholders to provide collective insights into the structure and behavior of a system through facilitated exercises and artifacts. GMB is often used to explicate the contrasting mental models of stakeholders.

Stakeholder An individual or group that has significant interest or influence over a policy problem.

System dynamics An analytic approach to problem definition and solution that focuses on endogenous variables linked through feedback, information and material delays, and non-linear relationships. The structure of these linkages determines the behavior of the modeled system.

Definition of the Subject

System dynamics is an approach to problem understanding and solution. It captures the complexity of real-world problems through the explication of feedback among endogenous variables. This feedback, and the delays that accompany it, often drive public sector programs towards unanticipated or unsatisfactory results. Through formal and informal modeling, System Dynamics-based analysis explicates and opens these feedback structures to discussion, debate and consensus building necessary for successful public sector policymaking.

Introduction

In the 50 years since its founding, System Dynamics has contributed to public policy thought in a number of areas. Major works, such as *Urban Dynamics* [35] and *Limits to Growth* [61] have sparked controversy and debate. Other works in the domains of military policy, illegal drugs, welfare reform, health care, international development, and education have provided deep insight into complex social problems. The perspective of System Dynamics, with its emphasis on feedback, changes over time, and the role of information delays, helps inform policy makers about the intended and unintended consequences of their choices. The System Dynamics method includes a problem-oriented focus and the accommodation of multiple stakeholders, both crucial to the development of sound policy. Through the use of formal simulation, decision makers may also use System Dynamics models to consider the effects of their choices on short- and long-term outcomes. We illustrate this process with real life examples, followed by a review of the features of System Dynamics as they relate to public policy issues. We then describe the conjunction of System Dynamics and Group Model Building as a mechanism for policy ideation and review. We identify some of the historical and current uses of System Dynamics in the public sector, and discuss techniques for evaluating its effects on policy and organizations.

Medical Malpractice: A System Dynamics and Public Policy Vignette

The year was 1987 and New York's medical malpractice insurance system was in a state of crisis. Fueled by unprecedented levels of litigation, total settlements were soaring as were the malpractice insurance rates charged to hospitals and physicians. Obstetricians stopped taking on new patients. Doctors threatened to or actually did leave the state. Commercial insurance carriers had stopped underwriting malpractice insurance policies, leaving state-sponsored risk pools as the only option. The Governor and the Legislature were under pressure to find a solution and to find it soon. At the center of this quandary was the state's Insurance Department, the agency responsible for regulating and setting rates for the state's insurance pools. The agency's head found himself in just the kind of media hot seat one seeks to avoid in the public service.

An in-house SWAT team of actuaries, lawyers, and analysts had been working to present a fiscally sound and politically viable set of options for the Agency to consider and recommend to the legislature. They had been working with a team of System Dynamics modelers to gain better understanding of the root causes of the crisis. Working as a group, they had laid out a whole-system view of the key forces driving malpractice premiums in New York State. Their simulation model, forged in the crucible of group consensus, portrayed the various options on a "level playing field," each option being analyzed using a consistent set of operating assumptions. One option stood out for its ability to offer immediate malpractice insurance premium relief, virtually insuring a rapid resolution to the current crisis. An actuarial restructuring of future liabilities arising from future possible lawsuits relieved immediate pressure on available reserve funds. Upward pressure on premiums would vanish; a showdown in the legislature would be averted. Obviously, the Commissioner was interested in this option–who would not be?

"But what happens in the later years, after our crisis is solved?" he asked. As the team pored over the simulation model, they found that today's solution sowed the seeds for tomorrow's problems. Ten, fifteen, or maybe more years into the future, the deferred liabilities piled up in the system creating a secondary crisis, quite literally a second crisis caused by the resolution of the first crisis.

"Take that option off the table – it creates an unacceptable future," was the Commissioner's snap judgment. At that moment a politically appointed official had summarily dismissed a viable and politically astute "silver bullet" cure to a current quandary because he was thinking dynamically, considering both short-term and long-term effects of policy.

The fascinating point of the medical malpractice vignette is that the option taken off the table was indeed, in the short run, a "silver bullet" to the immediate crisis. The System Dynamics model projected that the solution's unraveling would occur long after the present Commissioner's career was over, as well as after the elected life span of the Governor who had appointed him and the legislators whose votes would be needed to implement the solution. His decision did not define the current problem solely in terms of the current constellation of stakeholders at the negotiations, each with their particular interests and points of view. His dynamic thinking posed the current problem as the result of a system of forces that had accumulated in the past. Symmetrically, his dynamic thinking looked ahead in an attempt to forecast what would be the future dynamic consequences of each option. Might today's solution become tomorrow's problem?

This way of thinking supported by System Dynamics modeling invites speculation about long-run versus short-run effects. It sensitizes policy makers to the pressure of future possible stakeholders, especially future generations who may come to bear the burden of our current decisions. It draws attention into the past seeking causes that may be buried at far spatial and temporal distances from current symptoms within the system. It seeks to understand the natural reaction time of the system, the period during which problems emerge and hence over which they need to be solved. System Dynamics-based analysis in the public sector draws analytic attention away from the riveting logic of the annual or biannual budget cycle, often focusing on options that will play themselves out years after current elected officials have left office. Such work is hard to do, but critical if one wants to think in systems terms.

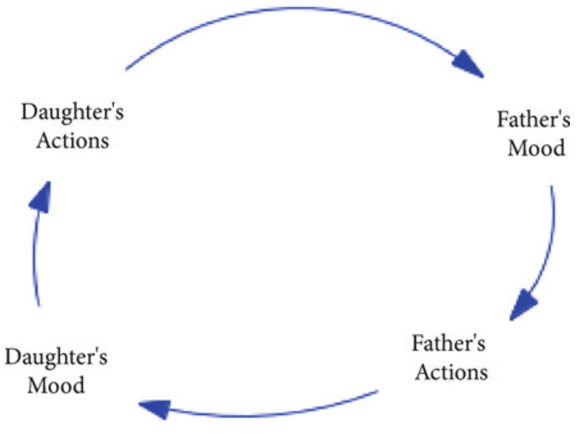
What Is System Dynamics Modeling?

While other papers in this series may provide a more expanded answer to this basic question, it may be useful to begin this discussion of System Dynamics and public policy with a brief description of what System Dynamics is.

System Dynamics is an approach to policy analysis and design that applies to problems arising in complex social, managerial, economic, or ecological systems [31,33,74,95]. System Dynamics models are built around a particular problem. The problem defines the relevant factors and key variables to be included in the analysis. This represents the model's boundary, which may cross departmental or organizational boundaries. One of the unique advantages of using System Dynamics models to study public policy problems is that assumptions from a variety of stakeholders are explicitly stated, can be tested through simulation, and can be examined in context.

System Dynamics models rely on three sources of information: numerical data, the written database (reports, operations manuals, published works, etc.), and the expert knowledge of key participants in the system [36]. The numerical database of most organizations is very small, the written database is larger, and the expert knowledge of key participants is vast. System Dynamicists rely on all three sources, with particular attention paid to the expert knowledge of key participants because it is only through such expert knowledge that we have any knowledge of the structure of the system. The explicit capturing of accumulated experience from multiple stakeholders in the model is one of the major differences between System Dynamics models and other simulation paradigms. An understanding of the long term effects of increased vigilance on the crime rate in a community needs to account for the reaction of courts, prisons, and rehabilitation agencies pressed to manage a larger population. This knowledge is spread across experts in several fields, and is not likely to be found in any single computer database. Rather, insight requires a process that makes these factors visible and explicit. For public sector problems, in particular, this approach helps move conflict out of the realm of inter-organizational conflict and towards a problem-solving focus.

Through the use of available data and by using the verbal descriptions of experts to develop mathematical relationships between variables, we expose new concepts and/or previously unknown but significant variables. System Dynamics models are appropriate to problems that arise in closed-loop systems, in which conditions are converted into information that is observed and acted upon, changing conditions that influence future decisions [69].



Public Policy, System Dynamics Applications to, Figure 1
Closed loop diagram of fathers and daughters

This idea of a "closed loop" or "endogenous" point of view on a system is really important to all good System Dynamics models. A simple example drawn from everyday life may help better to understand what an endogenous (versus exogenous) point of view means. If a father believes that his teenage daughter is always doing things to annoy him and put him in a bad mood, then he has an exogenous or "open loop" view of his own mood because he is seeing his mood as being controlled by forces outside of or exogenous to his own actions. However, if the father sees that his daughter and her moods are reacting to his own actions and moods while in turn his daughter's actions shape and define his moods, then this father has an endogenous point of view on his own mood. He understands how his mood is linked in a closed loop with another member of his family. Of course, the father with an endogenous view will be in a better position to more fully understand family dynamics and take actions that can prevent bad moods from spreading within the family.

Using an SD Model to Develop a Theory

A System Dynamics model represents a theory about a particular problem. Since models in the social sciences represent a theory, the most we can hope for from all these models, mental or formal, is that they be useful [94]. System Dynamics models are useful because the mathematical underpinning needed for computer simulation requires that the theory be precise. The process of combining numerical data, written data, and the knowledge of experts in mathematical form can identify inconsistencies about how we think the system is structured and how it behaves over time [38].

In policymaking it is often easy and convenient to blame other stakeholders for the problem state. Often, though, the structure of the system creates the problem by, for example, shifting resources to the wrong recipient or by inclusion of policies that intervene in politically visible but ineffective ways. The use of inclusive SD models educates us by identifying these inconsistencies through an iterative process involving hypotheses about system structure and tests of system behavior. Simulation allows us to see how the complex interactions we have identified work when they are all active at the same time. Furthermore, we can test a variety of policies quickly to see how they play out in the long run. The final result is a model that represents our most insightful and tested theory about the endogenous sources of problem behavior.

Behavior over Time Versus Forecasts

People who take a systems view of policy problems know that behavior generated by complex organizations cannot be well understood by examining the parts. By taking this holistic view, System Dynamicists capture time delays, amplification, and information distortion as they exist in organizations. By developing computer simulation models that incorporate information feedback, systems modelers seek to understand the internal policies and decisions, and the external dynamic phenomena that combine to generate the problems observed. They seek to predict dynamic implications of policy, not forecast the values of quantities at a given time in the future.

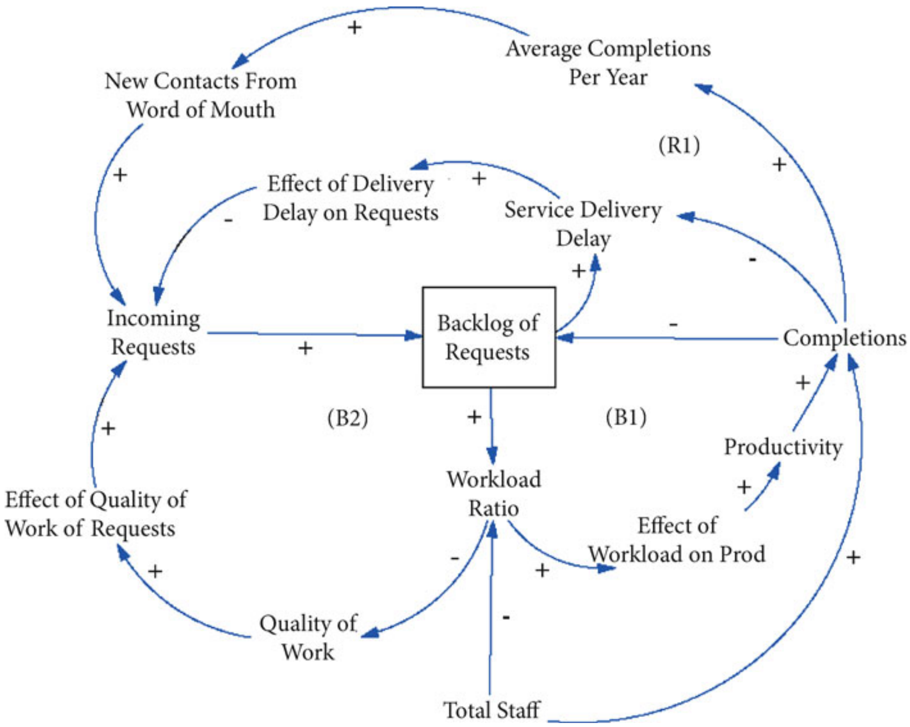
System Dynamics models are tools that examine the behavior of key variables over time. Historical data and performance goals provide baselines for determining whether a particular policy generates behavior of key variables that is better or worse, when compared to the baseline or other policies. Furthermore, models incorporating rich feedback structure often highlight circumstances where the forces governing a system may change in a radical fashion. For example, in early phases of its growth a town in an arid region may be driven by a need to attract new jobs to support its population. At some future point in time, the very fact of successful growth may lead to a water shortage. Now the search for more water, not more jobs, may be what controls growth in the system. Richardson [69] has identified such phenomena as shifts in loop dominance that provide endogenous explanations for specific outcomes. Simulation allows us to compress time [95] so that many different policies can be tested, the outcomes explained, and the causes that generate a specific outcome can be examined by knowledgeable people working in the system, before policies are actually implemented.

Excellent short descriptions of System Dynamics methodology are found in Richardson [69,70] and Barlas [9]. Furthermore, Forrester's [33] detailed explanation of the field in *Industrial Dynamics* is still relevant, and Richardson and Pugh [74], Roberts et al. [78], Coyle [22], Ford [31], Maani and Cavana [53], Morecroft [65] and Sterman [95] are books that describe the field and provide tools, techniques and modeling examples suitable for the novice as well as for experienced System Dynamics modelers.

An Application of System Dynamics – The Governor's Office of Regulatory Assistance (GORA) Example

When applied to public policy problems, the "nuts and bolts" of this System Dynamics process consist of identifying the problem, examining the behavior of key variables over time, creating a visualization of the feedback structure of the causes of the problem, and developing a formal simulation model. A second case illustration may assist in understanding the process. The New York State Governor's Office of Regulatory Assistance (GORA) is a governmental agency whose mission it is to provide information about government rules and regulations to entrepreneurs who seek to start up new businesses in the state. The case was described by Andersen et al. [7] and is often used as a teaching case introducing System Dynamics to public managers.

Figure 2 below illustrates three key feedback loops that contribute both to the growth and eventual collapse of citizen service requests at GORA. The reinforcing feedback loop labeled "R1" illustrates how successful completion of citizen orders creates new contacts from word-of-mouth by satisfied citizens which in turn leads to more requests for service coming into the agency. If only this loop were working, a self-reinforcing process would lead to continuing expansion of citizen requests for services at GORA. The balancing loop labeled "B2" provides a balancing effect. As workers within the agency get more and more work to complete, the workload within the agency goes up with one effect being a possible drop in the quality in the work completed. Over time, loop B2 tells a story of how an increased workload can lead to a lower quality of work, with the effect of that lower quality being fewer incoming requests in the future. So over time, too many incoming requests set off a process that limits future requests by driving down quality. Many public managers who have worked with the GORA model find these two simple feedback loops to be realistic and powerful explanations of many of the problems that their agencies face on a day-to-day basis. The full GORA model has many other


Average Completions Per Year
New Contacts From Word of Mouth
(R1)
Effect of Delivery Delay on Requests
Service Delivery Delay
Incoming Requests
Backlog of Requests
Completions
(B2)
(B1)
Productivity
Workload Ratio
Effect of Workload on Prod
Effect of Quality of Work of Requests
Quality of Work
Total Staff
+
-

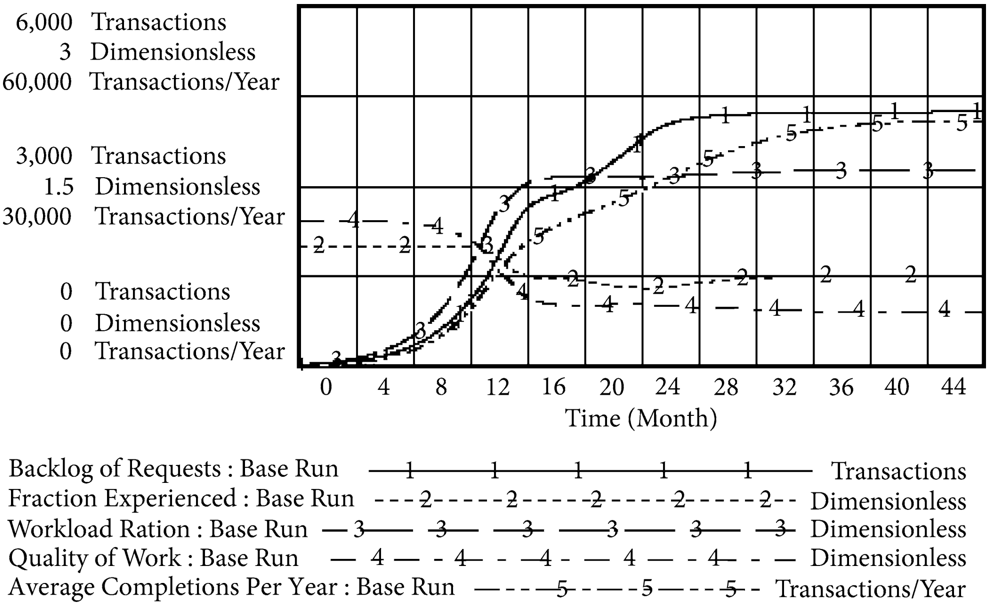
Public Policy, System Dynamics Applications to, Figure 2
Key feedback loops in a simulation of workflow in the Governor's Office of Regulatory Administration (GORA)

feedback loops and active variables not shown in the aggregated Fig. 2.

Once all of the variables have been represented by mathematical equations, a computer simulation is able to recreate an over time trajectory possible future values for all of the variables in the model. Figure 3 shows a graph over time of simulated data for key indicators in the GORA case study. The simulation begins when GORA comes into existence to provide services to the public and runs for 48 months. Initially, there is adequate staff and the amount of work to do is low, so the Workload Ratio, shown as part of loops B1 and B2 in the previous figure, is very low. With a low Workload Ratio GORA employees are able to devote additional time to each task they perform and the Quality of Work[1] is thus relatively high. The Backlog of Requests and the Average Completions Per Year begin at 0 and then increase and level off over time to approximately 4,500 and 41,000 respectively. The Fraction Experienced Staff measures what portion of the overall workers are experienced and hence more efficient at doing their jobs. As shown in Fig. 3, the Fraction Experienced begins at 1 and then falls and increases slightly to .75 indicating that GORA is having a harder time retaining experienced staff and is experiencing higher employee turnover. (The full GORA model has a theory of employee burnout and turnover not shown in Fig. 2.)

The combination of the visualization in Fig. 2 with a formal model capable of generating the dynamic output shown in Fig. 3 illustrates the power of System Dynamics modeling for public policy issues. Linking behavior and structure helps stakeholders understand why the behavior of key variables unfolds over time as it does. In the GORA case, the program is initially successful as staff are experienced, are not overworked, and the quality of the services they provide is high. As clients receive services the R1 feedback loop is dominant and this attracts new clients to GORA. However, at the end of the first year the number of clients requesting services begins to exceed the ability of GORA staff to provide the requested services in a timely manner. The Workload Ratio increases, employees are very busy, the Quality of Work falls, and the B2 feedback loop works to limit the number of people seek-

[1] The Workload Ratio and Quality of Work are normalized variables. This means that they are measured against some predetermined standard. Therefore, when these two variables are equal to 1 they are operating in the desired state. Depending on the definition of the variable, values below or above 1 indicate when they are operating in a desired or undesired state. For example, Quality of Work above 1 indicates that quality is high, relative to the predetermined normal. However, Quality of Work below 1 indicates an undesirable state.



Public Policy, System Dynamics Applications to, Figure 3
Simulated performance of key variables in the GORA case study

ing services. Furthermore, people are waiting longer to receive services and some are discouraged from seeking services due to the delay. The initial success of the program cannot be sustained and the program settles down into an unsatisfactory situation where the Workload Ratio is high, Quality of Work is low, clients are waiting longer for services and staff turnover is high as indicated by the Fraction Experienced.

The model tells a story of high performance expectations, initial success and later reversal, all explained endogenously. Creating and examining the simulation helps managers consider possible problems before they occur – before staff are overtaxed, before turnover climbs, and before the agency has fallen behind. Having a model to consider compresses time and provides the opportunity for a priori analysis. Finally, having a good model can provide managers with a test bed for asking "what if" questions, allowing public managers to spend simulated dollars and make simulated errors all the while learning how to design better public policies at relatively low cost and without real (only simulated) risk.

How Is System Dynamics Used to Support Public Policy and Management?

The Medical Malpractice vignette that opened this chapter involving the New York State Commissioner of Insurance is more fully documented by Reagan-Cirincione et al. [68] and is one of the first published examples of the results of a team of government executives working in a face-to-face group model building session to create a System Dynamics model to support critical policy decisions facing the group. The combined group modeling and simulation approach had a number of positive effects on the policy process. Those positive effects are:

Make Mental Models of Key Players Explicit

When the Commissioner drew together his team, the members of this group held different pieces of information and expertise. Much of the most important information was held in the minds, in the mental models, of the Commissioner's staff, and not in data tabulations. The System Dynamics modeling process made it possible for managers to explicitly represent and manipulate their shared mental models in the form of a System Dynamics simulation model. This process of sharing and aligning mental models, as done during a System Dynamics modeling intervention, is an important aspect of a "learning organization" as emphasized by Senge [87].

Create a Formal and Explicit Theory of the Public Policy Situation Under Discussion

The formal model of malpractice insurance contained an explicit and unambiguous theory of how the medical malpractice system in New York State functions. The shared mental models of the client team implied such a formal and model-based theory, but the requirements of creating a running simulation forced the group to be much more explicit and clear about their joint thinking. As the modeling team worked with the group, a shared consensus about how the whole medical malpractice system worked was cast, first into a causal-loop diagram, and later into the equations of the formal simulation model [74,95].

Document all Key Parameters and Numbers Supporting the Policy Debate

In addition to creating a formal and explicit theory, the System Dynamics model was able to integrate explicit data and professional experience available to the Department of Insurance. Recording the assumptions of the model in a clear and concise way makes possible review and examination by those not part of the model's development. Capturing these insights and their derivation provides face validity to the model's constructs.

Building confidence in the utility of a System Dynamics model for use in solving a public policy problem involves a series of rigorous tests that probe how the model behaves over time as well as how available data, both numerical and tacit structural knowledge, have been integrated and used in the model. Forrester and Senge [39] detail 17 tests for building confidence in a System Dynamics model. Sterman [95] identifies 12 model tests, the purpose or goal of each test, and the steps that modelers should follow in undertaking those tests. Furthermore, Sterman [95] also lists questions that model consumers should ask in order to generate confidence in a model. This is particularly important for public policy issues where the ultimate goal or outcome for different stakeholders may be shared, but underlying assumptions of the stakeholders may be different.

Create a Formal Model that Stimulates and Answers Key "what if" Questions

Once the formal model was constructed, the Commissioner and his policy team were able to explore "what if" scenarios in a cost-free and risk-free manner. Significant cost overruns in a simulated environment do not drive up real tax rates, nor do they lead to an elected official being voted out of office, nor to an appointed official losing her job. Quite the contrary, a simulated cost overrun or a simulated failed program provides an opportunity to learn how better to implement or manage the program or policy (or to avoid trying to implement the policy). Public managers get to experiment quickly with new policies or programs in a risk-free simulated environment until they"get it right" in the simulated world. Only then should they take the risk of implementation in a high stakes policy environment.

Bringing a complex model to large groups sometimes requires the development of a more elaborate simulation, so that those who were not part of the initial analysis can also derive insight from its results. Iterative development and discussion provides an additional validation of the constructs and conclusions of the model, Zagonel et al. [108] have described a case where local managers responsible for implementing the 1996 federal welfare reform legislation used a simulation model to explore such "what if" futures before taking risks of actual implementation.

Public policy problems are complex, cross organizational boundaries, involve stakeholders with widely different perspectives, and evolve over time. Changes in police procedures and/or resources may have an effect on prison and parole populations many years into the future. Health care policies will determine how resources are allocated at local hospitals and the types of treatments that can be obtained. Immigration policies in one country may influence the incomes and jobs of people in a second country. Miyakawa [64] has pointed out that public policies are systemically interdependent. Solutions to one problem often create other problems. Increased enforcement of immigration along the U.S. borders has increased the workload of courts [26]. Besides being complex these examples also contain stakeholders with different sets of goals. In solving public policy problems, how diverse stakeholders work out their differences is a key component of successful policy solutions. System Dynamics modeling interventions, and in particular the techniques of group model building [2,6,72,98], provide a unique combination of tools and methods to promote shared understanding by key stakeholders within the system.

System Dynamics and Models: A Range of Analytic Scope and Products

In the malpractice insurance example, the Commissioner called his advisors into a room to explicitly engage in a group model building session. These formal group model-building sessions involve a specialized blend of projected computer support plus professional facilitation in a face-to-face meeting of public managers and policy analysts. Figure 4 is an illustration of a team of public managers working together in a group model building project. In this photograph, a facilitator is working on a hand drawn view of a simulation model's structure while projected views of computer output can be used to look at

Public Policy, System Dynamics Applications to, Figure 4
A team of public managers working together to build a System Dynamics model of welfare reform policies

first cut simulation runs or refined images of the model being built by the group. Of course, the key feature of this whole process is facilitated face-to-face conversations between the key stakeholders responsible for the policy decisions being made.

Richardson and Andersen [72], Andersen and Richardson [6], Vennix [98], and Luna-Reyes et al. [52] have provided detailed descriptions of how this kind of group model building process actually takes place. In addition to these group model building approaches, the System Dynamics literature describes five other ways that teams of modelers work with client groups. They are (1) the Reference Group approach [91], (2) the Strategic Forum [75], (3) The stepwise approach [104], (4) strategy dynamics [100,101,102], and (5) the "standard method" of Hines [67].

Some System Dynamics-oriented analyzes of public policies completed by groups of public managers and policy analysts stop short of building a formal simulation model. The models produced by Wolstenholme and Coyle [107], Cavana, Boyd and Taylor [14] and the system archetypes promoted by Senge [87] have described how these qualitative system mapping exercises, absent a formal running simulation model, can add significant value to a client group struggling with an important public policy problem. The absence of a formal simulation limits the results to a conceptual model, rather than a tool for systematic experimentation.

Finally, a number of public agencies and Non Governmental Organizations are joining their counterparts in the private sector by providing broad-based systems thinking training to their top leaders and administrative staff. A number of simulation-based management exercises such as the production-distribution game (also known as the "beer game") [93] and the People's Express Flight Simulator [92] have been developed and refined over time to support such training and professional development efforts. In addition, Cavana and Clifford [11] have used GMB to develop a formal model and flight simulator to examine the policy implications of an excise tax policy on tobacco smoking.

What Are the Arenas in Which System Dynamics Models Are Used?

The malpractice insurance vignette and the GORA example represented cases where a model was developed for a single problem within one agency. Naill [66] provides an example of how a sustained modeling capability can be installed within an agency to support a range of ongoing policy decisions (in this case the model was looking at transitional energy policies at the federal level). Barney [10] developed a class of System Dynamics simulation models to support economic development and planning in developing nations. Wolstenholme [105] reported on efforts to support health planning within the British Health Service.

Addressing a tactical problem within a single public sector agency, while quite common, is only one of the many types of decision arenas in which System Dynamics models can be and are used to support public policy. Indeed, how a model is used in a public policy debate is largely determined by the unique characteristics of the specific decision-making arena in which the model is to be used. Some of the more common examples follow.

Models Used to Support Inter-Agency and Inter-Governmental Collaborative Efforts

A quite different arena for the application of System Dynamics models to support the policy process occurs when an interagency or inter-governmental network of program managers must cooperate to meet a common mission. For example, Rohrbaugh [79] and Zagonel et al. [108] report a case where state and local officials from social services, labor, and health agencies combined their efforts with private and non-profit managers of day care services, health care services, and worker training and education services to plan for comprehensive reform of welfare policies in the late 1990s. These teams were seeking strategies to blend financial and program resources across a myriad of stovepipe regulations and reimbursement schemes to provide a seamless system of service to clients at the local level. To complete this task, they created a simulation model containing a wide range of system-level interactions and tested policies in that model to find out what blend of policies might work. Policy implementation followed this model-based and simulation-supported policy design.

Models Used to Support Expert Testimony in Courtroom Litigation

Cooper [20] presented one of the first published accounts of a System Dynamics model being used as a sort of expert witness in courtroom litigation. In the case he reported, Litton Industries was involved in a protracted lawsuit with the U. S. Navy concerning cost and time overruns in the construction of several naval warships. In a nutshell, the Navy contended that the cost overruns were due to actions taken (or not taken) by Litton Industries as primary contractor on the project and as such the Navy should not be responsible for covering cost overruns. Litton maintained that a significant number of change orders made by the Navy were the primary drivers of cost overruns and time delays. A simulation model was constructed of the shipbuilding process and the simulation model then built two simulated ships without any change orders. A second set of "what if" runs subsequently built the same ships except that the change orders from the Navy were included in the construction process. By running and re-running the model, the analysts were able to tease out what fractions of the cost overrun could reasonably be attributed to Litton and what fraction should be attributed to naval change orders. Managers at Litton Industries attribute their receipt of hundreds of millions dollars of court-sanctioned payments to the analysis supported by this System Dynamics simulation model. Ackermann, Eden and Williams [1] have used a similar approach involving soft systems approaches combined with a System Dynamics model in litigation over cost overruns in the channel tunnel project.

Models Used as Part of the Legislative Process

While System Dynamics models have been actively used to support agency-level decision making, inter-agency and inter-governmental task forces and planning, and even courtroom litigation, their use in direct support of legislative processes has a more uneven track record. For example, Ford [30] reports successes in using System Dynamics modeling to support regulatory rule making in the electric power industry, and Richardson and Lamitie [73] report on how System Dynamics modeling helped redefine a legislative agenda relating to the school aid formula in the U.S. state of Connecticut. However, Andersen [4] remains more pessimistic about the ability of System Dynamics models to directly support legislative decision making, especially when the decisions involve zero-sum tradeoffs in the allocation of resources (such as formula-driven aid involving local municipal or education formulas). This class of decisions appears to be dominated by short-term special interests. A longer-term dynamic view of such immediate resource allocation problems is less welcome. The pathway to affecting legislative decision making appears to be by working through and with public agencies, networks of providers, the courts, or even in some opinions, by directly influencing public opinion.

Models Used to Inform the Public and Support Public Debate

In addition to using System Dynamics modeling to support decision making in the executive, judicial, and legislative branches of government (often involving Non-Governmental Organizations and private sector support), a number of System Dynamics studies appeal directly to the public. These studies intend to affect public policy by shaping public opinion in the popular press and the policy debate. In the 1960s, Jay Forrester's *Urban Dynamics* [34] presented a System Dynamics model that looked at many of the problems facing urban America in the latter half of the 20th century. Several years later in response to an invitation from the Club of Rome, Forrester put together a study that led to the publication of *World Dynamics* [35], a highly aggregate System Dynamics model that laid out a feedback-oriented view of a hypothesized set of relationships between human activity on the planet, industrialization, and environmental degradation. Meadows et al. [61]

followed on this study with a widely hailed (and critiqued) System Dynamics simulation study embodied in the best-selling book, *Limits to Growth*. Translated into over 26 languages, this volume coalesced a wide range of public opinion leading to a number of pieces of environmental reform in the decade of the 1970s. The debate engendered by that volume continues even 30 years later [63]. Donella Meadows continued in this tradition of appealing directly to public opinion through her syndicated column, The Global Citizen, which was nominated for the Pulitzer Prize in 1991. The column presented a System Dynamics-based view of environment matters for many years (http://www.pcdf.org/meadows/).

What Are some of the Substantive Areas Where System Dynamics Has Been Applied?

The International System Dynamics Society (http://www.systemdynamics.org) maintains a comprehensive bibliography of over 8,000 scholarly books and articles documenting a wide variety of applications of System Dynamics modeling to applied problems in all sectors. MacDonald et al. [54] have created a bibliography extracted from this larger database that summarizes some of the major areas where System Dynamics modeling has been applied to public policy. Below, we summarize some of the substantive areas where System Dynamics has been applied, giving one or two sample illustrations for each area.

Health Care

System Dynamicists have been applying their tools to analyze health care issues at both the academic and practitioner level for many years. *The System Dynamics Review*, the official journal of the System Dynamics Society, devoted a special issue to health care in 1999 due to the importance of health care as a critical public policy issue high on the political agenda of many countries and as an area where much System Dynamics work has been performed. The extensive System Dynamics work performed in the health care area fell into three general categories: patient flow management, general health policy, and specific health problems.

The patient flow management category is exemplified by the work of Wolstenholme [106], Lane and Rosenhead [48], and Van Ackere and Smith [97]. The articles written by these authors focused on issues and policies relating to patient flows in countries where health care service is universal.

The general health policy category is rather broad in that these articles covered policy and decision making from the micro level [96] to the macro level [88]. There were also many articles that showed how the process of modeling resulted in better understanding of the problem and issues facing health care providers and policy makers [12].

The last category dealt with specific health-related problems such as the spread of AIDS [43,76], smoking [42], and malaria control [29], as well as many other health-related conditions.

Education

The education articles touched on various topics relating to education ranging from using System Dynamics in the classroom as a student-centered teaching method to models that dealt with resource allocations in higher education. Nevertheless, many of the articles fell into five categories that could be labeled management case studies or flight simulators, teaching technology, research, teaching, and education policy.

The management case study and flight simulator articles are best exemplified by Sterman's [93] article describing the Beer Game and Graham, Morecroft et al. [41] article on “Model Supported Case Studies for Management Education.” The emphasis of these works is on the use of case studies in higher education, with the addition of games or computer simulations. This is related to the teaching technology category in that both emphasize using System Dynamics models/tools to promote learning. However, the teaching technology category of articles stresses the introduction of computer technology, specifically System Dynamics computer technology, into the classroom. Steed [90] has written an article that discusses the cognitive processes involved while using Stella to build models, while Waggoner [99] examined new technologies versus traditional teaching approaches.

In addition to teaching technology are articles that focus on teaching. The teaching category is very broad in that it encompasses teaching System Dynamics in K-12 and higher education as subject matter [37,77] as well as ways to integrate research into the higher-education classroom [71]. System Dynamics models are also used to introduce advanced mathematical concepts through simulation and visualization, rather than through equations [27,28]. In addition, lesson plans for the classroom are also part of this thread [44]. The Creative Learning Exchange (http://www.clexchange.com) provides a central repository of lessons and models useful for pre-college study of System Dynamics, including a selfstudy roadmap to System Dynamics principles [23].

There are also a number of articles that pertain to resource allocation [15] at the state level for K-12 schools along with articles that deal with resource-allocation decisions in higher education [32,40]. Saeed [83] and Mashayekhi [56] cover issues relating to higher education policy in developing countries.

The last education category involved research issues around education. These articles examined whether the System Dynamics methodology and simulation-based education approaches improved learning [24,47,55].

Defense

System Dynamics modeling work around the military has focused on manpower issues, resource allocation decisions, decision making and conflict. Coyle [21] developed a System Dynamics model to examine policies and scenarios involved in sending aircraft carriers against land-based targets. Wils, Kamiya et al. [103] have modeled internal conflicts as a result of outbreaks of conflict over allocation and competition of scarce resources. The manpower articles focused on recruitment and retention policies in the armed forces and are represented in articles by Lopez and Watson [51], Andersen and Emmerichs [5], Clark [18], Clark, McCullough et al. [19] and Cavana et al. [14]. The resource allocation category deals with issues of money and materials, as opposed to manpower, and is represented by Clark [16,17]. Decision making in military affairs from a System Dynamics perspective is represented in the article by Bakken and Gilljam [8].

Environment

The System Dynamics applications dealing with environmental resource issues can be traced back to when the techniques developed in *Industrial Dynamics* were beginning to be applied to other fields. The publication of Forrester's *World Dynamics* in 1971 and the follow-up study *Limits to Growth* [61,62,63] used System Dynamics methodology to address the problem of continued population increases on industrial capital, food production, resource consumption and pollution. Furthermore, specific studies dealing with DDT, mercury and eutrophication of lakes were part of the Meadows et al. [59] project and appeared as stand-alone journal articles prior to being published as a collection in Meadows and Meadows [60].

The environmental applications of System Dynamics have moved on since that time. Recent work has combined environmental and climate issues with economic concerns thorough simulation experiments [25] as well as stakeholder participation in environmental issues [89]. In 2004, the *System Dynamics Review* ran a special issue dedicated to environmental issues. Cavana and Ford [13] were the editors and did a review of the System Dynamics bibliography in 2004, identifying 635 citations with the key words "environmental" or "resource." Cavana and Ford broke the 635 citations into 11 categories they identified as resources, energy, environmental, population, water, sustainable, natural resources, forest, ecology, agriculture, pollution, fish, waste, earth, climate and wildlife.

General Public Policy

The System Dynamics field first addressed the issue of public policy with Forrester's *Urban Dynamics* [34] and the follow-up work contained in *Readings in Urban Dynamics* [58] and Alfeld and Graham's *Introduction to Urban Dynamics* [3]. The field then branched out into the previously mentioned *World Dynamics* and the follow-up studies related to that work. Moreover, the application of System Dynamics to general public policy issues began to spread into areas as diverse as drug policy [50], and the causes of patient dropout from mental health programs [49], to ongoing work by Saeed [82,84,85] on development issues in emerging economies. More recently, Saysel et al. [86] have examined water scarcity issues in agricultural areas, Mashayekhi [57] reports on the impact on public finance of oil exports in countries that export oil and Jones et al. [46] cover the issues of sustainability of forests when no single entity has direct control.

This brief review of the literature where System Dynamics modeling has been used to address public policy issues indicates that the field is making inroads at the micro level (within government agencies) and at the macro level (between government agencies). Furthermore, work has been performed at the international level and at what could truly be termed the global level with models addressing public policy issues aimed at climate change.

Evaluating the Effectiveness of System Dynamics Models in Supporting the Public Policy Process

System Dynamics modeling is a promising technology for policy development. But does it really work? Over the past several decades, a minor cottage industry has emerged that purports to document the successes (and a few failures) of System Dynamics models by reporting on case studies. These case studies report on successful applications and sometimes analyze weaknesses, making suggestions for improvement in future practice. Rouwette et al. [81] have compiled a meta-analysis of 107 such case-based stories.

However, as compelling as such case stories may be, case studies are a famously biased and unsystematic way to evaluate effectiveness. Presumably, failed cases will not be commonly reported in the literature. In addition, such a research approach illustrates in almost textbook fashion the full litany of both internal and external threats to validity, making such cases an interesting but unscientific compilation of war stories. Attempts to study live management teams in naturally occurring decision situations can have high external validity but almost always lack internal controls necessary to create scientifically sound insights.

Huz et al. [45] created an experimental design to test for the effectiveness of a controlled series of group-based System Dynamics cases in the public sector. They used a wide battery of pre- and post survey, interview, archival, administrative data, and qualitative observation techniques to evaluate eight carefully matched interventions. All eight interventions dealt with the integration of mental health and vocational rehabilitation services at the county level. Four of the eight interventions contained System Dynamics modeling sessions and four did not. These controlled interventions were designed to get at the impact of System Dynamics modeling on the public policy process.

Overall, Huz et al. [45] envisioned that change could take place in nine domains measured across three separate levels of analysis as illustrated in Table 1 below.

Using the battery of pre- and post test instruments, Huz found important and statistically significant results in eight of the nine domains measured. The exceptions were in domain 9 where they did not measure client outcomes, in domain 5 where "participants were not significantly more aligned in their perceptions on strategies for changes" (but were more aligned in goals), and in domain 7 where "no significant change was found with respect to structural conditions within the network" (but two other dimensions of organizational relationships did change).

Public Policy, System Dynamics Applications to, Table 1
Domains of measurement and evaluation used to assess impact of systems-dynamics interventions (see p. 151 in [45])

Level I	**Reflections of the modeling team**
Domain 1	Modeling team's assessment of the intervention
Level II	**Participant self-reports of the intervention**
Domain 2	Participants' perceptions of the intervention
Domain 3	Shifts in participants' goal structures
Domain 4	Shifts in participants' change strategies
Domain 5	Alignment of participant mental models
Domain 6	Shifts in understanding how the system functions
Level III	**Measurable system change and "bottom line" results**
Domain 7	Shifts in network of agencies that support services integration
Domain 8	Changes in system-wide policies and procedures
Domain 9	Changes in outcomes for clients

In their meta-analysis of 107 case studies of System Dynamics applications, Rouwette et al. [81] coded case studies with respect to eleven classes of outcomes, sorted into individual level, group level, and organizational level. The 107 cases were dominated by for-profit examples with 65 such cases appearing in the literature followed by 21 cases in the non-profit sector, 18 cases in governmental settings, and three cases in mixed settings. While recognizing possible high levels of bias in reported cases as well as difficulties in coding across cases and a high number of missing categories, they found high percentages of positive outcomes along all 11 dimensions of analysis. For each separate dimension, they analyzed between 13 and 101 cases with the fraction of positive outcomes for each dimension ranging from a low of 83% to several dimensions where 100% of the cases reporting on a dimension found positive results. At the individual level, they coded for overall positive reactions to the work, insight gained from the work, and some level of individual commitment to the results emerging from the study. At the group level, they coded for increased levels of communication, the emergence of shared language, and increases in consensus or mental model alignment. Organizational level outcomes included implementation of system level change. With respect to this important overall indicator they "found 84 projects focused on implementation, which suggests that in half (42) of the relevant cases changes are implemented. More than half (24) of these changes led to positive results"(see p. 20 in [81]).

Rouwette [80] followed this meta-analysis with a detailed statistical analysis of a series of System Dynamics-based interventions held mostly in governmental settings in the Netherlands. He was able to estimate a statistical model that demonstrated how System Dynamics group model building sessions moved both individuals and groups from beliefs to intentions to act, and ultimately on to behavioral change.

In sum, attempts to evaluate System Dynamics interventions in live settings continue to be plagued by methodological problems that researchers have struggled to overcome with a number of innovative designs. What is emerging from this body of study is a mixed, "good news and bad news" picture. All studies that take into account a reasonable sample of field studies show some successes and some failures. About one-quarter to one-half of the System Dynamics studies investigated showed low impact on decision making. On the other hand, roughly half of the studies have led to system-level implemented change with approximately half of the implemented studies being associated with positive measures of success.

Summary: System Dynamics – A Powerful Tool to Support Public Policy

While recognizing and respecting the difficulties of scientific evaluation of System Dynamics studies in the public sector, we remain relentlessly optimistic about the method's utility as a policy design and problem-solving tool. Our glass is half (or even three-quarters) full. A method that can deliver high decision impact up to three-quarters of the time and implement results in up to half of the cases examined (and in a compressed time frame) is a dramatic improvement over alternative approaches that can struggle for months or even years without coming to closure on important policy directions.

System Dynamics-based modeling efforts are effective because they join the minds of public managers and policy makers in an emergent dialog that relies on formal modeling to integrate data, other empirical insights, and mental models into the policy process. Policy making begins with the pre-existing mental models and policy stories that managers bring with them into the room. Policy consensus and direction emerge from a process that combines social facilitation with technical modeling and analysis. The method blends dialog with data. It begins with an emergent discussion and ends with an analytic framework that moves from "what is" baseline knowledge to informed "what if" insights about future policy directions.

In sum, we believe that a number of the process features related to building System Dynamics models to solve public policy problems contribute to their appeal for frontline managers:

- **Engagement** Key managers can be in the room as the model is evolving, and their own expertise and insights drive all aspect of the analysis.
- **Mental models** The model-building process uses the language and concepts that managers bring to the room with them, making explicit the assumptions and causal mental models managers use to make their decisions.
- **Complexity** The resulting nonlinear simulation models lead to insights about how system structure influences system behavior, revealing understandable but initially counterintuitive tendencies like policy resistance or "worse before better" behavior.
- **Alignment** The modeling process benefits from diverse, sometimes competing points of view as stakeholders can have a chance to wrestle with causal assumptions in a group context. Often these discussions realign thinking and are among the most valuable portions of the overall modeling effort.
- **Refutability** The resulting formal model yields testable propositions, enabling managers to see how well their implicit theories match available data about overall system performance.
- **Empowerment** Using the model managers can see how actions under their control can change the future of the system.

System Dynamics modeling projects merge managers' causal and structural thinking with the available data, drawing upon expert judgment to fill in the gaps concerning possible futures. The resulting simulation models provide powerful tools to develop a shared understanding and to ground what-if thinking.

Future Directions

While the field of System Dynamics has reached its half-centenary in 2007, its influence on public policy continues to grow. Many of the problems defined by the earliest writers in the field continue to challenge us today. The growing literature base of environmental, social, and education policy is evidence of continued interest in the systems perspective. In addition, System Dynamics modeling is growing in popularity for defense analysis, computer security and infrastructure planning, and emergency management. These areas have the characteristic problems of complexity and uncertainty that require the integration of multiple perspectives and tacit knowledge that this method supports. Researchers and practitioners will continue to be attracted to the open nature of System Dynamics models as a vehicle for consensus and experimentation.

We anticipate that the tool base for developing and distributing System Dynamics models and insights will also grow. Graphical and multimedia-based simulations are growing in popularity, making it possible to build clearer models and disseminate insights easily. In addition, the development of materials for school-age learners to consider a systems perspective to social problems gives us optimism for the future of the field, as well as for future policy.

Bibliography

Primary Literature

1. Ackermann F, Eden C, Williams T (1997) Modeling for Litigation: Mixing Qualitative and Quantitative Approaches. Interfaces 27(2):48–65
2. Akkermans H, Vennix J (1997) Clients' Opinions on Group Model-Building: An Exploratory Study. Syst Dyn Rev 13(1):3–31

3. Alfeld L, Graham A (1976) Introduction to Urban Dynamics. Wright-Allen Press, Cambridge
4. Andersen D (1990) Analyzing Who Gains and Who Loses: The Case of School Finance Reform in New York State. Syst Dyn Rev 6(1):21–43
5. Andersen D, Emmerichs R (1982) Analyzing US Military Retirement Policies. Simulation 39(5):151–158
6. Andersen D, Richardson GP (1997) Scripts For Group Model Building. Syst Dyn Rev 13(2):107–129
7. Andersen D, Bryson J, Richardson GP, Ackermann F, Eden C, Finn C (2006) Integrating Modes of Systems Thinking into Strategic Planning Education and Practice: The Thinking Persons' Institute Approach. J Public Aff Educ 12(3):265–293
8. Bakken B, Gilljam M (2003) Dynamic Intuition in Military Command and Control: Why it is Important, and How It Should be Developed. Cogn Technol Work (5):197–205
9. Barlas Y (2002) System Dynamics: Systemic Feedback Modeling for Policy Analysis. In: Knowledge for Sustainable Development, an Insight into the Encyclopedia of Life Support Systems, vol 1. UNESCO-EOLSS, Oxford, pp 1131–1175
10. Barney G (1982) The Global 2000 Report to the President: Entering the Twenty-First Century. Penguin, New York
11. Cavana RY, Clifford L (2006) Demonstrating the utility for system dynamics for public policy analysis in New Zealand: the case for excise tax policy on tobacco. Syst Dyn Rev 22(4):321–348
12. Cavana RY, Davies P et al (1999) Drivers of Quality in Health Services: Different Worldviews of Clinicians and Policy Managers Revealed. Syst Dyn Rev 15(3):331–340
13. Cavana RY, Ford A (2004) Environmental and Resource Systems: Editor's Introduction. Syst Dyn Rev 20(2):89–98
14. Cavana RY, Boyd D, Taylor R (2007) A Systems Thinking Study of Retnetion and Recruitment Issues for the New Zealand Army Electronic Technician Trade Group. Syst Res Behav Sci 24(2):201–216
15. Chen F, Andersen D et al (1981) A Preliminary System Dynamics Model of the Allocation of State Aid to Education. Dynamica 7(1):2–13
16. Clark R (1981) Readiness as a Residual of Resource Allocation Decisions. Def Manag J 1:20–24
17. Clark R (1987) Defense Budget Instability and Weapon System Acquisition. Public Budg Financ 7(2):24–36
18. Clark R (1993) The Dynamics of US Force Reduction and Reconstitution. Def Anal 9(1):51–68
19. Clark T, McCullough B et al (1980) A Conceptual Model of the Effects of Department of Defense Realignments. Behav Sci 25(2):149–160
20. Cooper K (1980) Naval Ship Production: A Claim Settled and Framework Built. Interfaces 10(6):20
21. Coyle RG (1992) A System Dynamics Model of Aircraft Carrier Survivability. Syst Dyn Rev 8(3):193–213
22. Coyle RG (1996) System Dynamics Modelling: A Practical Approach. Chapman and Hall, London
23. Creative Learning Exchange (2000) Road Maps: A Guide to Learning System Dynamics. Available at http://sysdyn.clexchange.org/road-maps/home.html
24. Davidsen P (1996) Educational Features of the System Dynamics Approach to Modelling and Learning. J Struct Learn 12(4):269–290
25. Fiddaman TS (2002) Exploring Policy Options with a Behavioral Climate-Economy Model. Syst Dyn Rev 18(2):243–267
26. Finely B (2006) Migrant Cases Burden System: Rise in Deportations Floods Detention Centers, Courts. Denver Post, 10/2/06: http://www.denverpost.com/immigration/ci_4428563
27. Fisher D (2001) Lessons in Mathematics: A Dynamic Approach. iSee Systems, Lebanon
28. Fisher D (2004) Modeling Dynamic Systems: Lessons for a First Course. iSee Systems, Lebanon
29. Flessa S (1999) Decision Support for Malaria-Control Programmes – a System Dynamics Model. Health Care Manag Sci 2(3):181–91
30. Ford A (1997) System Dynamics and the Electric Power Industry. Syst Dyn Rev 13(1):57–85
31. Ford A (1999) Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems. Island Press, Washington, DC
32. Forsyth B, Hirsch G et al (1976) Projecting a Teaching Hospital's Future Utilization: A Dynamic Simulation Approach. J Med Educ 51(11):937–9
33. Forrester J (1961) Industrial Dynamics. Pegasus Communications, Cambridge
34. Forrester J (1969) Urban Dynamics. Pegasus Communications, Waltham
35. Forrester J (1971) World Dynamics. Pegasus Communications, Waltham
36. Forrester J (1980) Information Sources for Modeling the National Economy. J Am Stat Assoc 75(371):555–566
37. Forrester J (1993) System Dynamics as an Organizing Framework for Pre-College Education. Syst Dyn Rev 9(2):183–194
38. Forrester J (1994) Policies, Decisions, and Information Sources for Modeling. Modeling for Learning Organizations. In: Morecroft J, Sterman J (eds) Productivity Press. Portland, OR, pp 51–84
39. Forrester J, Senge P (1980) Tests for Building Confidence in System Dynamics Models. In: Legasto Jr. AA et al (eds) System Dynamics. North-Holland, New York, 14, pp 209–228
40. Galbraith P (1989) Mathematics Education and the Future: A Long Wave View of Change. Learn Math 8(3):27–33
41. Graham A, Morecroft J et al (1992) Model Supported Case Studies for Management Education. Eur J Oper Res 59(1):151–166
42. Homer J, Roberts E et al (1982) A Systems View of the Smoking Problem. Int J Biomed Comput 13 69–86
43. Homer J, St Clair C (1991) A Model of HIV Transmission Through Needle Sharing. A Model Useful in Analyzing Public Policies, Such as a Needle Cleaning Campaign. Interfaces 21(3):26–29
44. Hopkins P (1992) Simulating Hamlet in the Classroom. Syst Dyn Rev 8(1):91–100
45. Huz S, Andersen D, Richardson GP, Boothroyd R (1997) A Framework for Evaluating Systems Thinking Interventions: An Experimental Approach to Mental Health System Change. Syst Dyn Rev 13(2)149–169
46. Jones A, Seville D et al (2002) Resource Sustainability in Commodity Systems: The Sawmill Industry in the Northern Forest. Syst Dyn Rev 18(2):171–204
47. Keys B, Wolfe J (1996) The Role of Management Games and Simulations in Education Research. J Manag 16(2):307–336
48. Lane DC, Monefeldt C, Rosenhead JV (1998) Emergency – But No Accident – A System Dynamics Study of an accident and emergency department. OR Insight 11(4):2–10

49. Levin G, Roberts E (eds) (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
50. Levin G, Hirsch G, Roberts E (1975) The Persistent Poppy: A Computer Aided Search for Heroin Policy. Ballinger, Cambridge
51. Lopez T, Watson J Jr (1979) A System Dynamics Simulation Model of the U. S. Marine Corps Manpower System. Dynamica 5(2):57–78
52. Luna-Reyes L, Martinez-Moyano I, Pardo T, Creswell A, Richardson GP, Andersen D (2007) Anatomy of a Group Model Building Intervention: Building Dynamic Theory from Case Study Research. Syst Dyn Rev 22(4):291–320
53. Maani KE, Cavana RY (2007) Systems Thinking, System Dynamics: Managing Change and Complexity. Pearson Education (NZ) Ltd, Auckland
54. MacDonald R et al (2007) System Dynamics Public Policy Literature. Syst Dyn Soc, http://www.systemdynamics.org/short_bibliography.htm
55. Mandinach E, Cline H (1993) Systems, Science, and Schools. Syst Dyn Rev 9(2):195–206
56. Mashayekhi A (1977) Economic Planning and Growth of Education in Developing Countries. Simulation 29(6):189–197
57. Mashayekhi A (1998) Public Finance, Oil Revenue Expenditure and Economic Performance: A Comparative Study of Four Countries. Syst Dyn Rev 14(2–3):189–219
58. Mass N (ed) (1974) Readings in Urban Dynamics. Wright-Allen Press, Cambridge
59. Meadows D, Beherens W III, Meadows D, Nail R, Randers J, Zahn E (ed) (1974) Dynamics of Growth in a Finite World. Pegasus Communications, Waltham
60. Meadows D, Meadows D (1977) Towards Global Equilibrium: Collected Papers. MIT Press, Cambridge
61. Meadows D, Meadows D, Randers J (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
62. Meadows D, Meadows D, Randers J (1992) Beyond the Limits. Chelsea Green Publishing Company, Post Mills
63. Meadows D, Randers J, Meadows D (2004) Limits to Growth: The 30-Year Update. Chelsea Green Publishing Company, White River Junction
64. Miyakawa T (ed) (1999) The Science of Public Policy: Essential Readings in Policy Sciences 1. Routledge, London
65. Morecroft J (2007) Strategic Modelling and Business Dynamics: A Feedback Systems Approach. Wiley, West Sussex
66. Naill R (1977) Managing the Energy Transition. Ballinger Publishing Co, Cambridge
67. Otto P, Struben J (2004) Gloucester Fishery: Insights from a Group Modeling Intervention. Syst Dyn Rev 20(4):287–312
68. Reagan-Cirincione P, Shuman S, Richardson GP, Dorf S (1991) Decision modeling: Tools for Strategic Thinking. Interfaces 21(6):52–65
69. Richardson GP (1991) System Dynamics: Simulation for Policy Analysis from a Feedback Perspective. In: Fishwick P, Luker P (eds) Qualitative Simulation Modeling and Analysis. Springer, New York
70. Richardson GP (1996) System Dynamics. In: Gass S, Harris C (eds) Encyclopedia of Operations Research and Management Science. Kluwer Academic Publishers, Norwell
71. Richardson G, Andersen D (1979) Teaching for Research in System Dynamics. Dynamica 5(3)
72. Richardson G, Andersen D (1995) Teamwork in Group Model Building. Syst Dyn Rev 11(2):113–137
73. Richardson G, Lamitie R (1989) Improving Connecticut School Aid: A Case Study with Model-Based Policy Analysis. J Educ Financ 15(2):169–188
74. Richardson G, Pugh J (1981) Introduction to System Dynamics Modeling. Pegasus Communications, Waltham
75. Richmond B (1997) The Strategic Forum: Aligning Objectives Strategy and Process. Syst Dyn Rev 13(2):131–148
76. Roberts C, Dangerfield B (1990) Modelling the Epidemiological Consequences of HIV Infection and AIDS: a Contribution from Operational Research. J Oper Res Soc 41(4):273–289
77. Roberts N (1983) An Introductory Curriculum in System Dynamics. Dynamica 9(1):40–42
78. Roberts N, Andersen DF, Deal RM, Grant MS, Schaffer WA (1983) Introduction to Computer Simulation: a System Dynamics Modeling Approach. Addison Wesley, Reading
79. Rohrbaugh J (2000) The Use of System Dynamics in Decision Conferencing: Implementing Welfare Reform in New York State. In: Garson G (ed) Handbook of Public Information Systems. Marcel Dekker, New York, pp 521–533
80. Rouwette E (2003) Group Model Building as Mutual Persuasion. Wolf Legal Publishers, Nijmegen
81. Rouwette E, Vennix J, Van Mullekom T (2002) Group Model Building Effectiveness: a Review of Assessment Studies. Syst Dyn Rev 18(1):5–45
82. Saeed K (1994) Development Planning and Policy Design: A System Dynamics Approach. Ashgate, Aldershot
83. Saeed K (1996) The Dynamics of Collegial Systems in the Developing Countries. High Educ Policy 9(1):75–86
84. Saeed K (1998) Towards Sustainable Development, 2nd Edition: Essays on System Analysis of National Policy. Ashgate, Aldershot
85. Saeed K (2003) Articulating Developmental Problems for Policy Intervention: A System Dynamics Modeling Approach. Simul Gaming 34(3):409–436
86. Saysel A, Barlas Y et al (2002) Environmental Sustainability in an Agricultural Development Project: A System Dynamics Approach. J Environ Manag 64(3):247–260
87. Senge P (1990) The Fifth Discipline: the Art and Practice of the Learning Organization. Doubleday/Currency, New York
88. Senge P, Asay D (1988) Rethinking the Healthcare System. Healthc Reform J 31(3):32–34, 44–45 5
89. Stave K (2002) Using SD to Improve Public Participation in Environment Decisions. Syst Dyn Rev 18(2):139–167
90. Steed M (1992) Stella, a Simulation Construction Kit: Cognitive Process and Educational Implications. J Comput Math Sci Teach 11(1):39–52
91. Stenberg L (1980) A Modeling Procedure for the Public Policy Scene. In: Randers J (ed) Elements of the System Dynamics Method. Pegasus Communications, Waltham, pp 257–288
92. Sterman J (1988) People Express Management Flight Simulator: Simulation Game, Briefing Book, and Simulator Guide. http://web.mit.edu/jsterman/www/SDG/MFS/PE.html
93. Sterman J (1992) Teaching Takes Off: Flight Simulators for Management Education. OR/MS Today (October):40–44
94. Sterman J (1996) A Skeptic's Guide to Computer Models. In: Richardson GP (ed) Modelling for Management. Dartmouth Publishing Company, Aldershot

95. Sterman J (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
96. Taylor K, Lane D (1998) Simulation Applied to Health Services: Opportunities for Applying the System Dynamics Approach. J Health Serv Res Policy 3(4):226–232
97. van Ackere A, Smith P (1999) Towards a Macro Model of National Health Service Waiting Lists. Syst Dyn Rev 15(3):225–253
98. Vennix J (1996) Group model building: Facilitating team learning using system dynamics. Wiley, Chichester
99. Waggoner M (1984) The New Technologies versus the Lecture Tradition in Higher Education: Is Change Possible? Educ Technol 24(3):7–13
100. Warren K (1999) Dynamics of Strategy. Bus Strateg Rev 10(3):1–16
101. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
102. Warren K (2005) Improving Strategic Management with the Fundamental Principles of System Dynamics. Syst Dyn Rev 21(4):329–350
103. Wils A, Kamiya M et al (1998) Threats to Sustainability: Simulating Conflict Within and Between Nations. Syst Dyn Rev 14(2–3):129–162
104. Wolstenholme E (1992) The Definition and Application of a Stepwise Approach to Model Conceptualization and Analysis. Eur J Oper Res 59(1):123–136
105. Wolstenholme E (1993) A Case Study in Community Care Using Systems Thinking. J Oper Res Soc 44(9):925–934
106. Wolstenholme E (1999) A Patient Flow Perspective of UK Health Services: Exploring the Case for New Intermediate Care Initiatives. Syst Dyn Rev 15(3):253–273
107. Wolstenholme E, Coyle RG (1983) The Development of System Dynamics as a Methodology for System Description and Qualitative Analysis. J Oper Res Soc 34(7):569–581
108. Zagonel A, Andersen D, Richardson GP, Rohrbaugh J (2004) Using Simulation Models to Address “What If” Questions about Welfare Reform. J Policy Anal Manag 23(4):890–901

Books and Reviews

Coyle RG (1998) The Practice of System Dynamics: Milestones, Lessons and Ideas from 30 Years of Experience. Syst Dyn Rev 14(4):343–365

Forrester J (1961) Industrial Dynamics. Pegasus Communications, Waltham

Maani KE, Cavana RY (2007) Systems Thinking, System Dynamics: Managing Change and Complexity. Pearson Education (NZ) Ltd, Auckland

MacDonald R (1998) Reducing Traffic Safety Deaths: A System Dynamics Perspective. In: 16th International Conference of the System Dynamics Society, Quebec ’98, System Dynamics Society Quebec City

Morecroft JDW, Sterman JD (eds) (1994) Modeling for Learning Organizations, System Dynamics Series. Productivity Press, Portland

Wolstenholme EF (1990) System Enquiry: A System Dynamics Approach. Wiley, Chichester

Scenario-Driven Planning with System Dynamics

NICHOLAS C. GEORGANTZAS
Fordham University Business Schools, New York, USA

Article Outline

Glossary

Mental model how one perceives cause and effect relations in a system, along with its boundary, i. e., exogenous variables, and the time horizon needed to articulate, formulate or frame a decision situation; one's implicit causal map of a system, sometimes linked to the reference performance scenarios it might produce.

Product either a physical good or an intangible service a firm delivers to its clients or customers.

Real option right and obligation to make a business decision, typically a tangible investment. The option to invest, for example, in a firm's store expansion. In contrast to financial 'call' and 'put' options, a *strategic real option* is not tradable. Any time it invests, a firm might be at once acquiring the *strategic real options* of expanding, downsizing or abandoning projects in future. Examples include research and development (abbreviated R&D), merger and acquisition (abbreviated M&A), licensing abroad and media options.

Scenario a postulated sequence or development of events trough time; via Latin *scena* '*scene*', from Greek *σκηνή*, *skēnē* '*tent, stage*'. In contrast to a forecast of what *will* happen in the future, a scenario shows what *might* happen. The term *scenario* must *not* be used loosely to mean situation. *Macro-environmental* as well as *industry-*, *task-* or *transactional-environmental* scenarios are merely inputs to the *strategic objectives* and *real options* a firm must subsequently explore through *strategic scenarios*, *computed* or simulated with an explicit, formal *system dynamics* (abbreviated SD) model of its strategic situation. *Computed strategic scenarios* create the multiple perspectives that strategic thinkers need to defeat the tyranny of dogmatism that often assails firms, governments and other social entities or organizations.

Scenario-driven planning (abbreviated SdP) to attain high performance through strategic flexibility, firms use the SdP *management technology* to create foresight and to anticipate the future with strategic real options, in situations where the business environment accelerates frequently and is highly complex or interdependent, thereby causing uncertainty.

Situation the set of circumstances in which a firm finds itself; its (strategic) state of affairs.

Strategic management process (abbreviated SMP) geared at detecting environmental threats and turning them into opportunities, it *proceeds from* a firm's mission, vision and environmental constraints *to* strategic goals and objectives *to* strategy design or formulation *to* strategy implementation or strategic action *to* evaluation and control *to* learning through feedback (background, Fig. 2).

SMP-1 environmental scanning monitors, evaluates and disseminates knowledge about a firm's internal and external environments to its people. The internal environment contains *s*trengths and *w*eaknesses within the firm; the external shows future *o*pportunities and *t*hreats (abbreviated SWOT).

SMP-2 mission a firm's purpose, *raison d'être* or reason for being.

SMP-3 objectives performance (*P*) goals that SMP often quantifies for some *P* metrics.

SMP-4 policy decision-making guidelines that link strategy design or formulation to action or implementation tactics.

SMP-5 strategy a comprehensive plan that shows how a firm might achieve its mission and objectives. The three strategy levels are: corporate, business and process or functional.

SMP-6 strategy design or formulation the interactive, as opposed to antagonistic, interplay of strategic content and process that creates flexible long-range plans to turn future environmental threats into opportunities; includes internal strengths and weaknesses as well as strategic mission and objectives, and policy guidelines.

SMP-7 strategic action or implementation the process by which strategies and policies are put into action through the development of programs, processes, budgets and procedures.

SMP-8 evaluation and control sub-process that monitors activities and performance, comparing actual results with desired performance.

SMP-9 learning through feedback occurs as knowledge about each SMP element enables improving previous SMP elements (background, Fig. 2).

System an organized group of interrelated components, elements or parts working together for a purpose; parts might be either goal seeking or purposeful.

System dynamics (abbreviated SD) a lucid modeling method born from the need to manage business performance through time. Thanks to Forrester [23], who discovered that all change propagates itself through stock and flow sequences, and user-friendly SD software (*iThink*®, *Vensim*®, etc.), SD models let managers see exactly how and why, like other biological and social organizations, business firms perform the way they do. Unlike other social sciences, SD shows exactly how *feedback loops*, i. e., circular cause and effect chains, each containing at least one time lag or delay, interact within a system to determine its performance through time.

Variable or metric something that changes either though time or among different entities at the same time. An *internal change lever* is a decision or policy variable that a strategy-design modeling, or client, team controls. An *external change trigger* is an environmental or policy variable that a strategy-design modeling team does not control. Both *trigger* and *lever* variables can initiate change and be either endogenous or exogenous to a model of a system.

However certain our expectation, the moment foreseen
may be unexpected when it arrives
—T.S. Eliot

Definition of the Subject

Many of us live and work in and about business ecosystems with complex structures and behaviors. Some realize that poor performance often results from our very own past actions or decisions, which come back to haunt us. So business leaders in diverse industries and firms, such as *Airbus, General Motors, Hewlett-Packard, Intel* and *Merck*, use scenario-driven planning (SdP) with system dynamics (SD) to help them identify, design and apply high-leverage, sustainable solutions to dynamically complex strategic-decision situations. One must know, for example, if the effect of an environmental change or strategic action gets magnified through time or is dampened and smoothed away. What may seem insignificant at first might cause major disruption in performance. SdP with SD shows the causal processes behind such dynamics, so firms can respond to mitigate impacts on performance.

Accelerating change and complexity in the global business environment make firms and other social organizations abandon their *inactive, reactive* and *preactive* modes [2]. SdP with SD turns them *proactive*, so they can translate anticipation into action. To properly transform anticipation into action, computed with SD models, 'strategic scenarios' must meet four conditions: *consistency, likelihood, relevance* and *transparency* [37]. Combining SdP with SD for that purpose, with other tools, like actor and stakeholder purposes, morphological methods or probability might help avoid entertainment and explore all possible scenarios. Indeed, SdP with SD

> "does not stand alone… modeling projects are part of a larger effort… modeling works best as a complement to other tools, not as a substitute" (see p. 80 in [75]).

SdP with SD is a systematic approach to a vital top-management job: leading today's firm in the rapidly changing and highly complex global environment. Anticipating a world where product life cycles, technology and the mix of collective- and competitive-strategy patterns change at an unprecedented rate is hard enough. Moving ahead of it might prove larger than the talent and resources now available in leading firms. SdP with SD leads to a decisive integration of strategy design and operations, with the dividing line much lower than at present. As mid-level managers take on more responsibility, senior executives become free to give more time and attention to economic conditions, product innovation and the changes needed to enhance creativity toward strategic flexibility [23].

It is perhaps its capacity to reintegrate strategy content and process that turns SdP with SD into a new paradigm for competitive advantage [42], and simulation modeling in general [28], into a critical fifth tool, in addition to the four tools used in science: observation, logical-mathematical analysis, hypothesis testing and experiment [77]. But full-fledged SD models also allow computing scenarios to assess possible implications of strategic situations. Strategic scenarios are not merely hypothesized plausible futures, but computed by simulating combined changes in strategy and in the business environment [32].

Computed scenarios help managers understand what they do not know, enabling strategy design and implementation through the coalignment of timely tactics to improve long-term performance. Through its judicious use of resources, scenario-driven planning with system dynamics makes the tactics required for implementation clear [27]. And because computed scenarios reveal the required coalignment of tactics through time, SdP with SD helps firms become flexible, dependable and efficient, and save time!

Everyone's mind sees differently, but if there is truth in the adage 'a picture is worth a thousand words', then the complex interrelations that SdP with SD unearth and show must be worth billions. In a world where strategic chitchat dominates, one can only hope that SdP with SD will play a central role in public and private dialogues about dynamically complex opportunities and threats.

We shape our buildings; thereafter,
our buildings shape us
—Winston Churchill

Introduction

Following on the heels of Ackoff and Emery [3] and Christensen [10], respectively, Gharajedaghi [35] and Raynor [64] show how strategies with the best chances for brilliant success expose firms to debilitating uncertainty. Firms fail as their recipes for success turn bad through time. Gharajedaghi [35] shows, for example, five strategy scenarios that convert success to failure. Each scenario plays a critically different role. Together, however, these scenarios form a dynamically complex system. Through time, as each scenario plays, it enables the context for the next:

1. *Noble ape* or *copycat* strategy imitates and replicates advantage. Also called 'shadow marketing', it lets shadowy copycats instantly *shadow market* product technology, often disruptively.
2. *Patchy* or *sluggish* strategy delays responses to new technology. When this *second* scenario plays, then patching up wastes time, enabling competitors to deliver new technology and to dominate markets. Worse, it causes costs to rise as it drives down product quality.
3. *Satisficing* or *suboptimal* strategy scenarios take many forms. One entails a false assumption: if a policy lever helps produce desired performance, then pulling or pushing on that lever will push performance further.
4. *Gambling* or *changing the game* strategy scenario transforms a strategic situation by playing the game successfully. While dealing with a challenge, firms gradually transform their strategic situation and change the basis for competition, so a whole new game and set of issues emerge. Success marked, for example, the beginning of the *information era*. But competitive advantage has already moved away from having access to information. In our *systems era* [2], creating new knowledge and generating insight is the new game [81].

Lastly, the cumulative effects from these four strategy scenarios trickle down to the:

5. *Archetypal swing* or *paradigm shift* scenario. Both learning and unlearning can cause archetypal swings and paradigm shifts to unfold through time intentionally [76]. These also occur unintentionally when conventional wisdom fails to explain patterns of events that challenge prevailing mental models. The lack of a convincing explanation creates a twilight zone where acceptable ideas are not competent and competent ideas are not acceptable.

Beliefs about the future drive strategies. But the future is unpredictable. Worse, success demands commitments that make it impossible to adapt to a future that turns out surprising. So, strategies with great success potential also bear high failure probabilities. Raynor [64] calls this the *strategy paradox*. Dissolving it requires turning environmental uncertainty into strategic flexibility. To make it so, Raynor urges managers to: anticipate multiple futures with scenarios, formulate optimal strategies for each future, accumulate *strategic real options* [5] and manage the select options portfolio.

SdP with SD helps managers who operate in an uncertain world question their assumptions about how the world works, so they can see it more clearly. To survive, the human mind overestimates small risks and underestimates large risks. Likewise, it is much more sensitive to losses than to gains. So the capability to leverage opportunities and to mitigate risk might have become an economic value driver.

The purpose of computing scenarios is to help managers alter their view of reality, to match it up more closely with reality as is and as it might become. To become a leader, a manager must define reality. The SdP with SD purpose is *not*, however, to paint a more accurate picture of tomorrow, but to improve the quality of decisions about the future. Raynor says that the requisite strategic flexibility, which SdP with SD creates:

> "is not a pastiche of existing approaches. Integrating these tools and grounding them in a validated theory of organizational hierarchy creates something that is quite different from any of these tools on its own, or in mere combination with the others" (see p. 13 in [64]).

Indeed, knowledge of common purposes and the acceptable means of achieving them form and hold together a purposeful hierarchical system. Its members know and share values embedded in their culture, which knits parts into a cohesive whole. And because each part has a lot to say about the whole, consensus is essential to SdP with SD for the co-alignment of diverse interests and purposes.

Ackoff and Emery [3], Gharajedaghi [35] and Nicolis [55] concur that purpose offers the lens one needs to see a firm as a multi-minded social net. A purposeful firm produces either the same result differently in the same environment or different results in the same or different environments. Choosing among strategic real options is necessary but insufficient for purposefulness. Firms that behave differently but show only one result per environment are goal seeking, not purposeful. Servomechanisms are goal seeking but people are purposeful. As a purposeful system, the firm is part of purposeful sub-systems, such as its *industry value chain* [61] and the society. And firms have purposeful people as members. The result is a dynamically interdependent, i. e., complex, hierarchical purposeful system.

A firm's value chain is, along with its primary and support activities, at once a member of at least one industry value chain and of the society or *macro-environment*. Industry analysis requires looking at value chains independently from the society [61]. But people, the society and firm and industry value chains are so interdependent, so interconnected, that an optimal solution might not exist for any of them independently of the others. SdP with SD helps firms co-align the 'plural rationality' of purposeful stakeholder groups with each other and that of the system as a whole.

Seeing strategic management as a *strategies and tactics net* [27] is in perfect syzygy with the *plural rationality* that SdP with SD accounts for among individuals, groups and organizations. Singer [73,74] contrasts monothematic conventional universes of traditional rationality with the multiverse-directed view of plural rationality. In counterpoint, Morecroft's [52] computed scenarios trace the dysfunctional interactions among sales objectives, overtime and sales force motivation to the intended, i. e., stated, singular rationality that drove action in a large sales organization.

Because their superordinate purpose is neither to compete nor to collaborate, but to develop new wealth-creating capabilities, in unique ways that serve both current and future stakeholder interests, customers and clients included [51], firms can benefit from the multiverse-directed view of strategic management as a net of strategies and tactics. SdP with SD helps firms break free from the tradeoffs tyranny of the mass-production era. Evidently, adherents to tradeoffs-free strategy like *Bell Atlantic*, *Daimler-Benz*, *Hallmark* and *Motorola* "can have it all" [60].

A firm must serve the purposes of its people as well as those of its environment, not as a mindless mechanical system, but as a living, purposeful, *knowledge-bonded* hierarchical system [3,35,55,81]. To clarify, a bike always yields to its rider, for example, regardless of the rider's desire; even if that entails running into a solid brick wall. Ouch**!** But riding a horse is an entirely different story. Horse and rider form a knowledge-bonded system: the horse must know the rider and the rider must know exactly how to lead the horse.

SdP with SD History: Always Back, Always in Style, Always Practical

Herman Kahn introduced scenarios to planning while at RAND Corporation in the 1950s [45]. Scenarios entered military strategy studies conducted for the US government. In the 1960s, Ozbekhan [58] used urban planning scenarios in Paris, France. Organization theorists and even novelists were quick to catch on. The meaning of scenarios became literary. Imaginative improvisation produced flickering apocalyptic predictions of strikingly optimistic and pessimistic futures. Political and marketing experts use scenarios today to jazz up visions of favorable and unfavorable futures.

Wack [78,79] asserts it was Royal Dutch Shell that came up with the idea of scenarios in the early 1970s. Godet [36] points to the French OTAM team as the first to use scenarios in a futures study by DATAR in 1971. Brauers and Weber [8] claim that Battelle's scenarios method [49] was originally a German approach. In connection with planning, however, most authors see scenario methods as typically American.

Indeed, during the 1970s, US researchers Olaf Helmer and Norman Dalkey developed scenario methods at RAND for eliciting and aggregating group judgments via Delphi and cross-impact matrices [4]. They extended cross impact analysis within statistical decision theory [39]. A synthesis of scenario methods began in the 1970s that draws together multiple views, including those of professional planners, analysts and line managers.

Ansoff [6] and other strategy theorists state that the 1970s witnessed the transformation of global markets. Today, changes in the external sociopolitical environment become pivotal in strategy making. Combined with the geographical expansion of markets, they increase the complexity of managerial work. As environmental challenges move progressively faster, they increase the likelihood of strategic surprises. So, strategic thinkers use scenarios to capture the nonlinearity of turbulent environments. Examples are Hax and Majluf [38] and, more clearly so, Porter [61] and Raynor [64]. They consider scenarios instrumental both in defining uncertainty and in anticipating environmental trends.

Huss and Honton [41] see scenarios emerge as a distinct field of study, a hybrid of a few disciplines. They identify multiple scenarios methods that fall into three major categories:

1. Intuitive logics [78,79], now practiced by *SRI International*,
2. Trend-impact analysis, practiced by the *Futures Group* and
3. Cross-impact analysis, practiced by the *Center for Futures Research* using INTERAX (Interactive Cross-Impact Simulation) and by Battelle using BASICS (BAttelle Scenario Inputs to Corporate Strategies).

Similarly, after joining Ozbekhan to advocate reference scenarios, Ackoff [2] distinguishes between:

1. Reference projections as piecemeal extrapolations of past trends and
2. The overall reference scenario that results from putting them together.

Based on Acar's [1] work under Ackoff, Georgantzas and Acar [32] explore these distinctions with a practical managerial technology: *comprehensive situation mapping* (CSM). CSM is simple enough for MBA students to master in their capstone Business Policy course. With the help of *Vensim® PLE* [18], CSM computes scenarios toward achieving a well-structured process of managing ill-structured strategic situations. In their introduction to SD, Georgantzas and Acar (see Chap. 10 in [32]) draw from the banquet talk that Jay Wright Forrester, Germeshausen Professor Emeritus, MIT, gave at the 1989 *International Conference of the System Dynamics Society*, in Germany, at the University of Stuttgart:

After attending the Engineering College, University of Nebraska, which included control dynamics at its core, Forrester went to MIT. There he worked for Gordon S. Brown, a pioneer in feedback control systems. During World War II, Brown and Forrester worked on servomechanisms for the control of radar antennas and gun mounts. This was research toward an extremely practical end, during which Forrester run literally from mathematical theory to the battlefield, aboard the US carrier *Lexington*.

After the war, Forrester worked on an analog aircraft flight simulator that could do little more than solve its own internal idiosyncrasies. So, Forrester invented *random-access magnetic storage* or *core memory*. His invention went into the heart of *Whirlwind*, a digital computer used for experimental development of military combat systems that eventually became the *s*emi*a*utomatic *g*round *e*nvironment (SAGE) air defense system for North America.

Alfred P. Sloan, the man who built *General Motors*, founded the *Sloan School of Management* in 1952. Forrester joined the school in 1956. Having spent fifteen years in the science and engineering side of MIT, he took the challenge of exploring what engineering could do for management.

One day, he found himself among students from *General Electric*. Their household appliance plants in Kentucky puzzled them: they would work with three or four shifts for some time and then, a few years later, with half the people laid off. Even if business cycles would explain fluctuating demand, that did not seem to be the entire reason. *GE*'s managers felt something was wrong.

After talking with them about hiring, firing and inventory policies, Forrester did some simulation on a paper pad. He started with columns for inventories, employees and customer orders. Given these metrics and *GE*'s policies, he could tell how many people would be hired or fired a week later. Each decision gave new conditions for employment, inventories and production. It became clear that wholly determined internally, the system had potential for oscillatory dynamics. Even with constant incoming orders, the policies caused employment instability. That longform simulation of *GE*'s inventory and workforce system marked the beginning of system dynamics [23,24,25,26].

SdP with SD Use and Roadmap

Scenarios mostly help forecast alternative futures but, as firms abandon traditional forecasting methods for interactive planning systems, line managers in diverse business areas adopt scenario-driven planning with system dynamics. Realizing that a tradeoffs-free strategy design requires insight about a firm's environment, both business and sociopolitical, to provide intelligence at *all* strategy levels, firms use SdP with SD to design *corporate, business* and *process* or *functional* strategies. SdP with SD is not a panacea and requires discipline, but has been successful in many settings. Its transdisciplinary nature helps multiple applications, namely capital budgeting, career planning, civil litigation [31], competitive analysis, crisis management, decision support systems (DSS), macroeconomic analysis, marketing, portfolio management and product development [65]. SdP with SD is a quest for managers who wish to be leaders, not just conciliators. They recognize that *logical incrementalism*, a piecemeal approach, is inadequate when the environment and their strategy change together.

Top management might see both divisional, i. e., business, and process or functional strategies as ways of implementing corporate strategy. But *active subsidiaries* [43,44] provide both strategic ideas and results to their parent enterprise. Drawing too stiff a line between the corporate office and its divisions might be

> "an unhealthy side effect of our collective obsession with generating returns. The frameworks for developing competitive strategy that have emerged over the last thirty years have given us unparalleled insight into how companies can succeed. And competitive strategy remains enormously important, but it should be the preserve of divisional management… corporate strategy should be focused on the management of strategic uncertainty" (see p. 11 in [64]).

Roadmap It is material to disconnect scenarios from unproductive guesswork and to anchor them to sound practices for strategy design. This guided tour through the fascinating but possibly intimidating jungle of scenario definitions shows what the future might hold for SdP with SD. Extensive literature, examples, practical guidelines and two real-life cases show how computed scenarios help manage uncertainty, that necessary disciple of our open market system. Unlike extrapolation techniques, SdP with SD encourages managers to think broadly about the future.

The above sections clarify the required context and provide a glossary. Conceptual confusion leads to language games at best and to operational confusion at worst [15]. SdP with SD helps firms avert both types of confusion. Instead of shifting their focus away from actuality and rationality, managers improve their insight about fundamental assumptions underlying changes in strategy. The mind-set of SdP with SD makes it specific enough to give practical guidance to those managing in the real world, both now and in the future.

The sections below look at *three* SdP with SD facets linked to strategy design and implementation. The *first* facet involves the business environment, the forces behind its texture and future's requisite uncertainty (Sect. "Environmental Turbulence and Future Uncertainty"). The *second* entails unearthing unstated assumptions about changes in the environment and in strategy, and about their potential combined effects on performance. The SdP with SD framework (Sect. "SdP with SD: The Modeling Process ≡ Strategic Situation Formulation") builds on existing scenario methods. Its integrative view delineates processes that enhance institutional learning, bolster productivity and improve performance through strategic flexibility. It shows why interest in computed scenarios is growing.

The *third* facet entails *computing* the combined or mixed effects on performance of changes both in the environment and in strategy. Even in mature economies, no matter how and how frequently said, decision makers often forget how the same action yields different results as the environment changes. The result is often disastrous. Conversely, the tight coupling between computed scenarios and strategic results can create new knowledge. Linking a mixed environmental and decision scenario in a one-to-one correspondence to a strategic result suits the normative inclination of strategic management, placing rationalistic inquiry at par with purely descriptive approaches in strategy research.

The unified treatment of SdP with SD and the strategy-making process grants a practical bonus, accounting for the entry's peculiar nature. It is not only a conceptual or idea contribution, but also an application-oriented entry. Sections "Case 1: Cyprus' Environment and Hotel Profitability" and "Case 2: A Japanese Chemicals Keiretsu (JCK) present two real-life cases of scenario-driven planning with system dynamics. Written with both the concrete and the abstract thinker in mind, the two cases show how firms and organizations build scenarios with a modest investment. SdP with SD provides an effective management technology that serves well those who adopt it. It saves them both time and energy.

Improvements in causal mapping [19,20], and SD modeling and analysis [50,57] contribute to the SdP with SD trend (Sect. "Future Directions"). Behavioral decision theory and cognitive science also help translate the knowledge of managers into SD models. The emphasis remains on small, transparent models of strategic situations and on dialogue between the managers' mental models and the computed scenarios [53].

All prognosticators are bloody fools
—Winston Churchill

Environmental Turbulence and Future Uncertainty

Environmental Turbulence

Abundant frameworks describe the business environment, but the one by Emery and Trist [22], which Duncan [17] abridged, has been guiding many a strategic thinker. It shows four business environments, each more complex and troublesome for the firm than the preceding one (Fig. 1a).

1. *Placid* or *independent-static environment*: infrequent changes are independent and randomly distributed, i. e., IID. Surprises are rare, but no new major opportunities to exploit either (*cell* 1, Fig. 1a).
2. *Placid-clustered* or *complex-static environment*: patterned changes make forecasting crucial. Comparable to the economist's idea of imperfect competition, this environment lets firms develop distinctive competencies to fit limited opportunities that lead to growth and bureaucracy (*cell* 2, Fig. 1a).
3. *Disturbed-reactive* or *independent-dynamic environment*: firms might influence patterned changes. Comparable to oligopoly in economics, this environment makes changes difficult to predict, so firms increase their operational flexibility through decentralization (*cell* 3, Fig. 1a).
4. *Turbulent field* or *complex-dynamic environment*: most frequent, changes are also complex, i. e., interdependent, originating both from autonomous shifts in the environment and from interdependence among firms and conglomerates. Social values accepted by members guide strategic response (*cell* 4, Fig. 1a).

Ansoff and McDonnell [7] extend the dichotomous environmental uncertainty perceptions by breaking turbulent environments (*cell* 4, Fig. 1a) into *discontinuous* and *surprising*. This is a step in the right direction, but not as helpful as a causal model specific to the system structure of a firm's strategic situation. Assuredly, 2 × 2 typologies help clarify exposition and are most frequent in the organization theory and strategy literatures. The mystical significance of duality affected even Leibniz, who associated one with God and zero with nothingness in the binary system. The generic solutions that dichotomies provide leave out the specifics that decision makers need. No matter what business they are in (Fig. 1b), managers cannot wait until a better theory comes along; they must act now.

It is worth noting that people often confuse the term ‘complex’ with ‘complicated’. Etymology shows that *complicated* uses the Latin ending *-plic*: *to fold*, but *complex* contains the Greek root *πλέξ*- ‘*plēx*-’: *to weave*. A complicated structure is thereby folded, with hidden facets stuffed into a small space (Fig. 1c). But a complex structure has interwoven parts with mutual interdependencies that cause dynamic complexity [46]. Remember: complex is the opposite of independent or untwined (Fig. 1a) and complicated is the opposite of simple (Fig. 1c).

Daft and Weick's [12] vista on firm *intrusiveness* and *environmental equivocality* is pertinent here. They see many events and trends in the environment as being inherently unclear. Managers discuss such events and trends, and form mental models and visions expressed in a fuzzy language and label system [80]. Within an *enactment* process, equivocality relates to managerial assumptions underlying the *analyzability* of the environment. A firm's *intrusiveness* determines how *active* or *passive* the firm is about environmental scanning. In this context, as the global environment gets turbulent, active firms and their subsidiaries construct SdP with SD models and compute scenarios to improve performance.

Managers of active firms combine knowledge acquisition with interpretations about the environment and their strategic situation. They reduce equivocality by assessing alternative futures through computed scenarios. In frequent meetings and debates, some by videoconferencing, managers use the dialectical inquiry process for *s*trategic *a*ssumption *s*urfacing and *t*esting (SAST), a vital strategic loop. Often ignored, the SAST loop gives active firms a strategic compass [47].

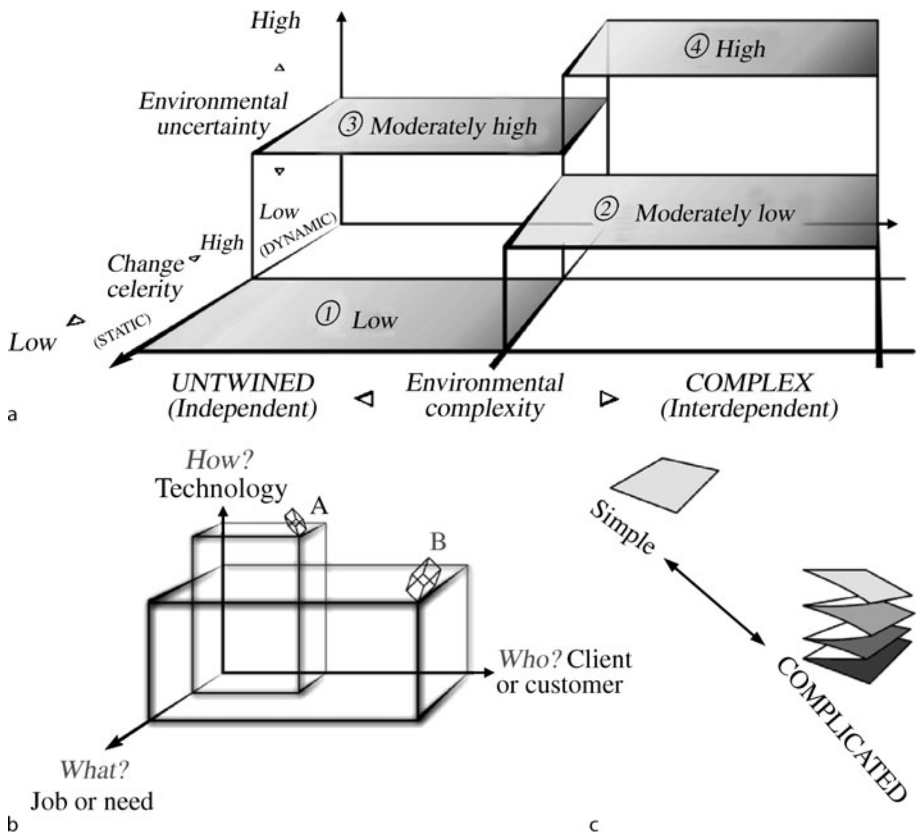
Conversely, passive firms do not actively seek knowledge but reduce equivocality through rules, procedures and regular reports: reams of laser-printed paper with little or no pertinent information. Managers in passive firms use the media to interpret environmental events and trends. They obtain insight from personal contacts with significant others in their environment. Data are personal and informal, obtained as the opportunity arises.

Future Uncertainty

“If we were omnipotent”, says Ackoff, then we could get “perfectly accurate forecasts” (see p. 60 in [2]). Thank God the future is unpredictable and we must yet create it. If it were not, then life would have been so boring! Here are some facts about straight forecasting:

1. Forecasts are seldom perfect, in fact, they are always wrong, so a useful forecasting model is one that minimizes error.
2. Forecasts always assume underlying stability in systems.
3. Product family and aggregated forecasts are always more accurate than single product forecasts, so the large numbers law applies.
4. In the short-term, managers can forecast but cannot act because time is too short; in the long term, they can act but cannot forecast.

To offset conundrum #4, SdP with SD juxtaposes the decomposition of performance dynamics into the growth and decline archetypes caused by *balancing* (–) and *reinforcing* (+) recursive causal-link chains or *feedback loops* [33,50]. A thermostat is a typical example of a goal-seeking feedback loop that causes either balancing growth



Scenario-Driven Planning with System Dynamics, Figure 1

a Environmental complexity and change celerity dimensions that cause perceived environmental uncertainty (adapted from [32]). b Scenario-driven planning with system dynamics helps with strategy-design fundamentals, such as, for example, defining a business along the requisite client-job-technology three-dimensional grid. c The simple-complicated dimension must *not* be confused with the environmental complexity dimension (adapted from [46])

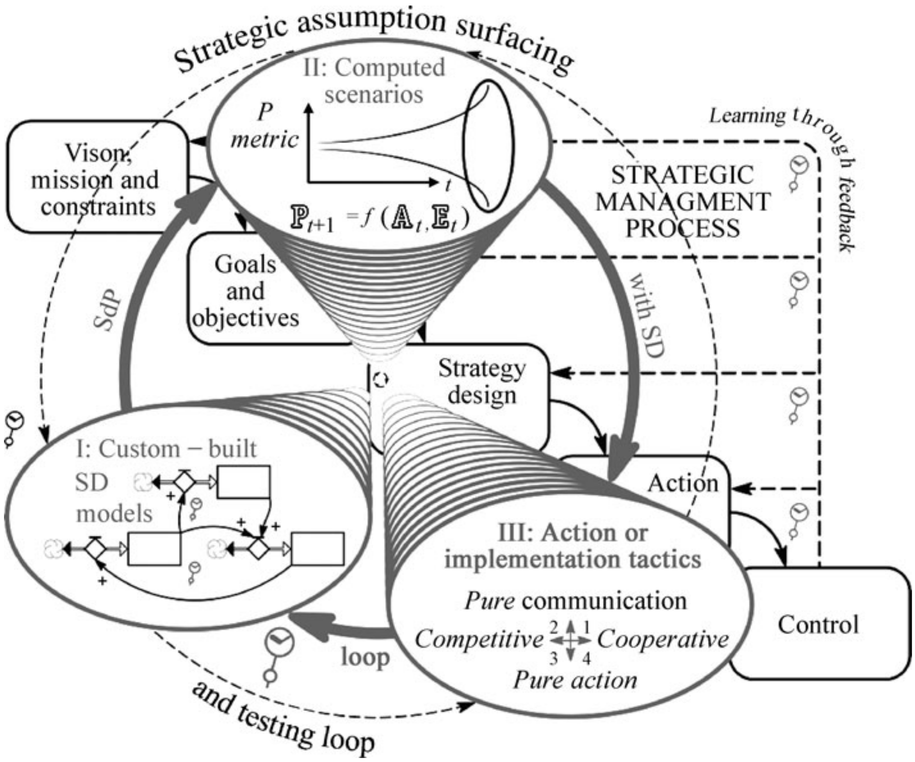
or decline. The gap between desired and room temperature causes action, which alters temperature with a time lag or delay. Temperature changes in turn close the gap between desired and room temperature.

Conversely, a typical loop that feeds on itself to cause either exponential growth or decline is that of an arms race. One side increases its arms. The other sides increase theirs. The first side then reacts by increasing its arms, and so on. Price wars between stores, promotional competition, shouting matches, one-upmanship and the wildcard interest rates of the late 1970s are good examples too. Escalation might persist until the system explodes or outside intervention occurs or one side quits, surrenders or goes out of business. In the case of wildcard interest rates, outside intervention by a regulatory agency can bring an end to irrationally escalating rates.

We've never been here before
—Peter Senge

SdP with SD: The Modeling Process ≡ Strategic Situation Formulation

The strategic management process (SMP, Fig. 2) starts with environmental scanning, in order to gauge environmental trends, opportunities and threats. Examples include increasing rivalry among existing competitors and Porter's [62] emphasis on the bargaining power of buyers and suppliers as well as on the threats of new entrants and substitutes. Even if some firms reduce environmental scanning to industry analysis in practice, changes in the environment beyond an industry's boundaries can determine what happens within the industry and its entry, exit and inertia barriers. Internal capability analysis comes next. It examines a firm's past actions and internal policy levers that can both propel and limit future actions. The integrative perspective of the SdP with SD framework on Fig. 2 delineates processes that enhance institutional


Strategic assumption surfacing
II: Computed scenarios
P metric
t
$\mathbb{P}_{t+1} = f(\mathbb{A}_t, \mathbb{E}_t)$
Vison, mission and constraints
Goals and objectives
SdP
with SD
STRATEGIC MANAGMENT PROCESS
Learning through feedback
Strategy design
Action
Control
I: Custom − built SD models
III: Action or implementation tactics
Pure communication
Competitive
Cooperative
Pure action
1
2
3
4
loop
and testing loop

Scenario-Driven Planning with System Dynamics, Figure 2

Cones of resolution show how scenario-driven planning with system dynamics enhances the strategy design component of the strategic management process (SMP; adapted from [32])

learning, bolster productivity and improve performance through strategic flexibility.

Strategy design begins by identifying variables pertinent to a firm's strategic situation, along with their interrelated causal links. Changes in these variables can have profound effects on performance. Some of the variables belong to a firm's external environment. Examples are emerging new markets, processes and products, government regulations and international interest and currency rates. Changes either in these or their interrelated causal links determine a firm's performance through time.

It is a manager's job to understand the causal links underlying a strategic situation. SdP with SD helps anticipate the effects of future changes triggered in the external environment. Other variables are within a firm's control. Pulling or pushing on these internal levers also affects performance. To evaluate a change in strategy, one must look at potential results along with changes in the environment, matching resource capabilities, stakeholder purposes, and organizational goals and objectives (Fig. 2).

Most variables interact. Often, the entire set of possible outcomes is obscure, difficult to imagine. But if managers oversimplify, then they end up ignoring the combined effects of chain reactions. Even well-intended rationality often leads to oversimplification, which causes cognitive biases (CBs) that mislead decision makers [21,70,72]. Conversely, computing mixed environmental and decision scenarios that link internal and external metrics can reveal unwarranted simplification.

SdP with SD integrates business intelligence with strategy design, not as a narrow specialty, but as an admission of limitations and environmental complexity. It also uses multiperspective dialectics, crucial for strategic assumption surfacing and testing (SAST). Crucial because the language and labels managers use to coordinate strategic real options are imprecise and fuzzy. Fuzzy language is not only adequate *initially* for managing interdependence-induced uncertainty but required [80]. Decision makers rely on it to overcome psychological barriers and Schwenk's [70] groups of CBs.

The best-case scenario for a passive firm is to activate modeling on Fig. 2, sometimes unknowingly. When its managers boot up, for example, electronic spreadsheets that contain inside-out causal models, with assumptions hidden deeply within many a formula. At bootup, only the numbers show. So passive-firm managers use electronic spreadsheets to laser-print matrices with comforting numbers. They

> "twiddle a few numbers and diligently sucker themselves into thinking that they're forecasting the future" [69].

And that is only when rapid changes in the environment force them to stop playing *blame the stakeholder*. They stop fighting the last war for a while, artfully name the situation a crisis, roll up their sleeves, and chat about and argue, but quickly agree on some arbitrary interpretation of the situation to generate strategic face-saving options. Miller and Friesen (see pp. 225–227 in [48]) show how for futile firms, rapid environmental changes lead to crisis-oriented decisions. Conversely, successful firms look far into the future as they counter environmental dynamism through strategy design with real options. Together, their options and interpretation of the environment, through the consensus that SdP with SD facilitates, enable a shared logic to emerge: a shared mental model that filters hidden spreadsheet patterns and heroic assumptions clean and clear.

Managers of active firms enter the SdP with SD loop of Fig. 2 both consciously and conscientiously. They activate strategic intelligence via computed scenarios and the SAST loop. Instead of twiddling spreadsheet numbers, *pro*active firm managers twiddle model assumptions. They stake, through SD model diagrams, their intuition about how they perceive the nature and structure of a strategic situation. Computed scenarios quantitatively assess their perceived implications. Having quantified the implications of shared visions and claims about the structure of the strategic situation, managers of active firms are likely to reduce uncertainty and equivocality. Now they can manage strategic interdependence. Because articulated perception is the starting point of all scenarios, computed scenarios give active firms a fair chance at becoming fast strategic learners.

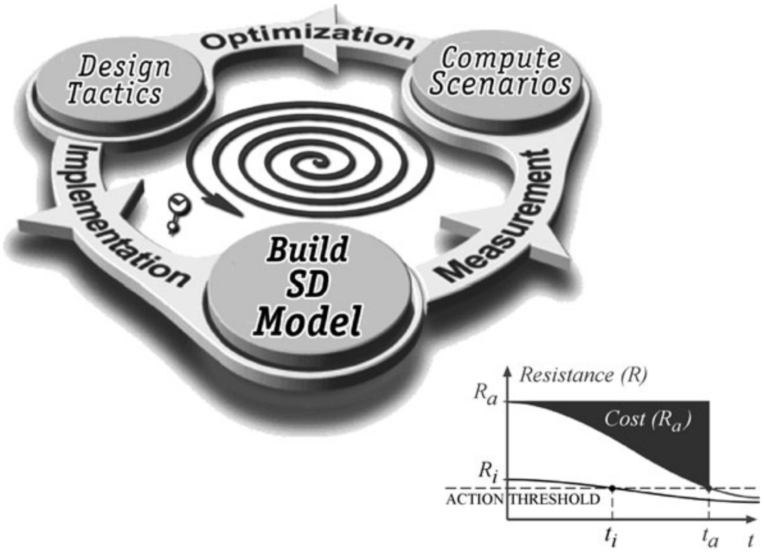
The design of action or implementation tactics requires detailing how, when and where a strategy goes into action. In addition to assuming the form of *pure communication* (III: 1 and 2, Fig. 2) or *pure action* (III: 3 and 4, Fig. 2), in a pragmatic sense, tactics can be either cooperative or competitive and defensive or offensive. Market location tactics, for example, can be either offensive, trying to rob market share from established competitors, or defensive, preventing competitors from stealing one's market share. An offensive tactic takes the form of frontal assault, flanking maneuver, encirclement, bypass attack or guerilla warfare. A defensive tactic might entail raising structural barriers, increasing expected retaliation or lowering the inducement for future attack. Conversely, cooperative tactics try to gain mutual advantage by working with rather than against others. Cooperative tactics take the form of alliances, joint ventures, licensing agreements, mutual service consortia and value-chain partnerships, the co-location of which often creates industrial districts [29].

The usual copycat strategy retort shows linear thinking at best and clumsy *benchmarking*, also known as shadow marketing, at worst. Its proponents assume performance can improve *incrementally*, with disconnected tactics alone, when strategy design is of primary concern. Piecemeal tactics can undermine strategy, but they are secondary. It might be possible to improve performance through efficient tactics, but is best to design strategies that expel counterproductive tactics. Counterproductive tactics examples are coercive moves that increase rivalry, without a real payoff, either direct or indirect, for the industry incumbent who initiates them. It is atypical of an industry or market leader to initiate such moves.

In strategy, superb action demands superior design. According to the design school, which Ansoff, Channon, McMillan, Porter, Thomas and others lead, logical incrementalism may help implementation, but becomes just another prescription for failure when the environment shifts. Through its judicious use of corporate resources, SdP with SD makes the tactics required for action clear. Also, it reveals their proper coalignment through time, so a firm can build strategic flexibility and save time!

The Modeling Process ≡ Strategic Situation Formulation

SdP with SD (Fig. 2) begins by modeling a business or 'social process' than a business or 'social system'. It is more productive to identify a *social process* first and then seek its causes than to slice a chunk of the real world and ask what dynamics it might generate. Distinguishing between a *social system* and a social process is roughly equivalent to distinguishing between a system's underlying causal structure and its dynamics. Randers (see p. 120 in [63]) defines a social system as a set of cause and effect relations. Its structure is a causal diagram or map of a real-world chunk. A social process is a behavior pattern of events evolving through time. The simulation results of SdP with SD models show such chains of events as they might occur in the real world. An example of a social system (structure) is the set of rules and practices that a firm might enact when dealing with changes in demand, along with the communication channels used for transmitting information and de-



Scenario-Driven Planning with System Dynamics, Figure 3
The recursive nature of the modeling process that scenario-driven planning with system dynamics entails a creates a sustainable, ever-expanding vortex of insight and wisdom, needed in strategic real-options valuation, and b saves both time and money as it renders negligible the cost of resistance (*R*) to change

cisions. A corresponding social process (dynamics) might be the stop-and-go pattern of capital investment caused by a conservative bias in a firm's culture.

In his model of a new, fast-growing product line, for example, Forrester [24] incorporates such a facet of corporate culture. Causing sales to stagnate, considerable back orders had to accumulate to justify expansion because the firm's president insisted on personally controlling all capital expenditures.

People often jump into describing system structure, perhaps because of its tangible nature as opposed to the elusive character of dynamics or social process fragments. Also, modelers present model structure first and then behavior. Ultimately, the goal in modeling a strategic situation is to link system structure and behavior. Yet, in the early stages of modeling is best to start with system dynamics and then seek underlying causes. Indeed, SD is particularly keen in understanding system performance, "not structure per se" (see p. 331 in [56]), in lieu of SD's core tenet that structure causes performance.

The modeling process itself is recursive in nature. The path from real-world events, trends and negligible externalities to an effective formal model usually resembles an expanding spiral (Fig. 3a). A useful model requires conceptualization; also focusing the modeling effort by establishing both the time horizon and the perspective from which to frame a decision situation. Typically, strategy-design models require a long-term horizon, over which computed scenarios assess the likely effects of changes both in strategy and in the environment.

Computer simulation is what makes SdP with SD models most useful. Qualitative cause and effect diagrams are too vague, tricky to simulate mentally. Produced through knowledge elicitation, their complexity vastly exceeds the human capacity to see their implications. Casting a chosen perspective into a formal SdP with SD model entails postulating a detailed structure; a diagramming description precise enough to propagate images of alternative futures, i. e., computed scenarios, "though not necessarily accurate" (see p. 118 in [63]). But the modeling process must never downplay the managers' mental database and its knowledge content. Useful models always draw on that mental database [24].

Following Morecroft [53], SdP with SD adopters might strive to replace the notion of modeling an objectively singular world *out there*, with the much softer approach of building formal models to improve managers' mental models. The expanding spiral of Fig. 3a shows that the insight required for decisive action increases as the quantity of information decreases, by orders of magnitude. The required quantification of the relations among variables pertinent to a strategic situation changes the character of the information content as one moves from mental to written to numerical data. Perceptibly, a few data remain, but much more pertinent to the nature and structure of the situation. Thanks to computed scenarios, clarity rules in the end. And, if the modeling process stays *i*nteractive (i), as opposed to *a*ntagonistic (a), then clarity means low resistance to change ($R_i < R_a$, Fig. 3b), which helps reach a firm's action/implementation threshold quickly ($t_i < t_a$, Fig. 3b). This is how SdP with SD users build strategic flexibility while they save both time and money!

Case 1: Cyprus' Environment and Hotel Profitability

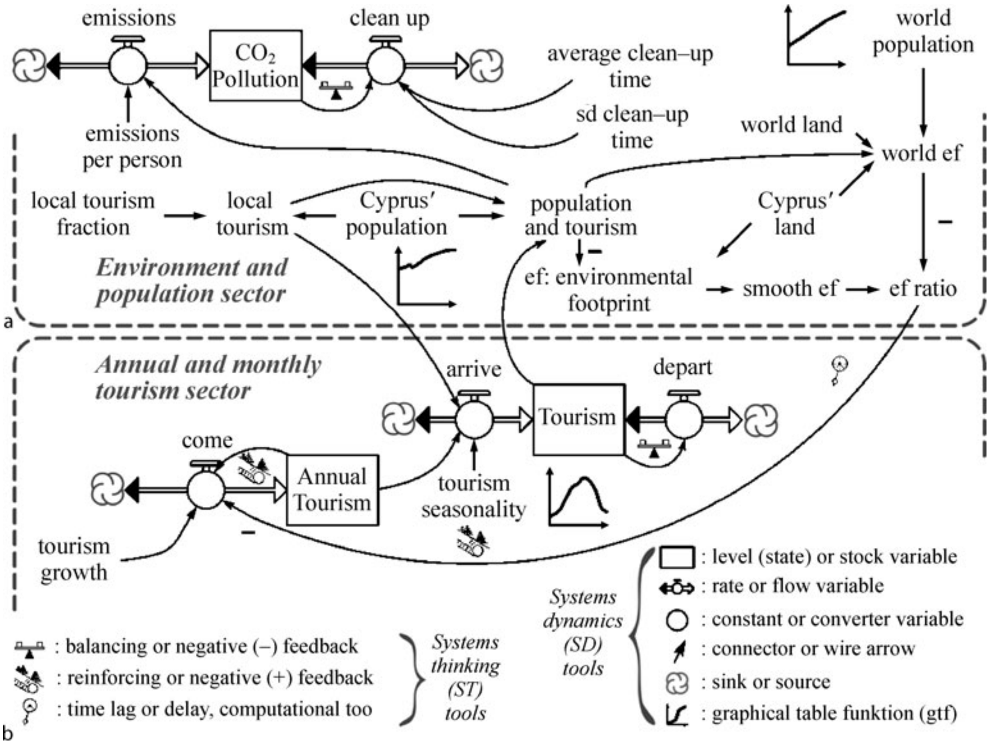
Cyprus' Hotel Association wished to test how Cyprus' year 2010 official tourism strategy might affect tourist arrivals, hotel bed capacity and profitability, and the island's environment [30]. Computed with a system dynamics simulation model, four tourism growth scenarios show what might happen to Cyprus' tourism over the next 40 years, along with its potential effects on the sustainability of Cyprus' environment and hotel profitability. Following is a partial description of the system dynamics model that precedes its dynamics.

Model Description (Case 1)

The SD model highlights member interactions along Cyprus' hotel value chain. The model incorporates a generic value-chain management structure that allows modeling customer-supplier value chains in business as well as in physical, biological and other social systems. Although the structure is generic, its situation specific parameters faithfully reproduce the dynamic behavior patterns seen in Cyprus' hotel value-chain processes, business rules and resources.

Cyprus' Environment, Population and Tourism Model Sectors Within Cyprus' environment and population sector (Fig. 4a), the carbon dioxide (CO_2) pollution stock is the accumulation of Cyprus' anthropogenic emissions less the Mediterranean Sea region's self clean-up rate. The clean-up rate that drains Cyprus' CO_2 pollution depends on the level of anthropogenic pollution itself as well as on the average clean-up time and its standard deviation (sd). Emissions that feed CO_2 pollution depend on Cyprus' population and tourism and on emissions per person [9].

In SD models, rectangles represent stocks, i. e., level or state variables that accumulate through time, e. g., the Tourism stock on Fig. 4b. The double-line, pipe-and-valve-like icons that fill and drain the stocks, often emanating from cloud-like sources and ebbing into cloud-like sinks, represent material flows that cause the


emissions
CO₂ Pollution
clean up
average clean–up time
sd clean–up time
world population
world land
world ef
emissions per person
local tourism fraction
local tourism
Cyprus' population
population and tourism
Cyprus' land
–
ef: environmental footprint
smooth ef
ef ratio
Environment and population sector
a
Annual and monthly tourism sector
arrive
Tourism
depart
come
Annual Tourism
tourism seasonality
tourism growth
–
: level (state) or stock variable
: rate or flow variable
: constant or converter variable
: connector or wire arrow
: sink or source
: graphical table funktion (gtf)
Systems dynamics (SD) tools
: balancing or negative (–) feedback
: reinforcing or negative (+) feedback
: time lag or delay, computational too
Systems thinking (ST) tools
b

Scenario-Driven Planning with System Dynamics, Figure 4
Cyprus' a environment and population, and b annual and monthly tourism model sectors (adapted from and extending [30])

Scenario-Driven Planning with System Dynamics, Table 1

Cyprus' environment and population (and local tourism) model sector (Fig. 4a) equations, with variable, constant parameter and unit definitions

Level or state variables **(stocks)**	***Eq.*** #
CO_2Pollution(t) = CO_2Pollution(t − dt) + (emissions − clean up) ∗ dt	(1.1)
INIT CO_2 Pollution = emissions (Based on 1995 gridded carbon dioxide anthropogenic emission data; unit: 1000 metric ton C per one degree latitude by one degree longitude grid cell)	(1.1.1)
Rate variables **(flows)**	
Emissions = emissions per person ∗ population and tourism (unit: 1000 metric tons C/month)	(1.2)
Cleanup = max(0, CO_2Pollution/average clean − up time) (unit: 1000 metric tons C/month)	(1.3)
Auxiliary variables and constants **(converters)**	
Average clean − up time = 1200 (Med Sea region average self clean-up time = 100 years; unit: months)	(1.4)
Cyprus' land = If (time $\le$ 168) then (9251 ∗ 247.1052) else ((9251 − 3355) ∗ 247.1052) (Cyprus' free land area; unit: acres; 1 km^2 = 247.1052 acres)	(1.5)
EF ratio = smooth EF/world EF (unit: unitless)	(1.6)
EF: environmental footprint = Cyprus' land/population and tourism (unit: acres/person)	(1.7)
Emissions per person = 1413.4/702000/12 (unit: anthropogenic emissions/person/month)	(1.8)
Local tourism = local tourism fraction ∗ Cyprus' population (unit: persons/month)	(1.9)
Local tourism fraction = 0.46 ∗ (0.61 + 0.08) (Percentages based on a 1995 study on domestic tourism; unit: unitless)	(1.10)
Population and tourism = Cyprus' population + Tourism − local tourism (Subtracts local tourists already included in Cyprus' population; unit: persons)	(1.11)
Sd clean − up time = 240 (clean-up time standard deviation = 20 years; unit: months)	(1.12)
Smooth EF = SMTH3 (EF: environmental footprint, 36) (Third-order exponential smooth of EF)	(1.13)
World EF = (world land − Cyprus' land)/(world population − population and tourism) (unit: acres/person)	(1.14)
World land = 36677577730.80 (unit: acres)	(1.15)
Cyprus' population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 493984), (1.00, 498898), (2.00, 496570), (3.00, 502001), (4.00, 505622), (5.00, 509329), (6.00, 512950), (7.00, 516743), (8.00, 520968), (9.00, 525364), (10.0, 529847), (11.0, 534330), (12.0, 539934), (13.0, 546486), (14.0, 552348), (15.0, 526313), (16.0, 516054), (17.0, 515881), (18.0, 518123), (19.0, 521657), (20.0, 526744), (21.0, 532692), (22.0, 538210), (23.0, 544675), (24.0, 551659), (25.0, 558038), (26.0, 560366), (27.0, 568469), (28.0, 572622), (29.0, 578394), (30.0, 587392), (31.0, 598217), (32.0, 609751), (33.0, 619658), (34.0, 626534), (35.0, 632082), (36.0, 636790), (37.0, 641169), (38.0, 645560), (39.0, 649759), (40.0, 653786), (41.0, 657686), (42.0, 661502), (43.0, 665246), (44.0, 668928), (45.0, 672554), (46.0, 676147), (47.0, 679730), (48.0, 683305), (49.0, 686870), (50.0, 690425), (51.0, 693975), (52.0, 697524), (53.0, 701056), (54.0, 704547), (55.0, 707970), (56.0, 711305), (57.0, 714535), (58.0, 717646), (59.0, 720613), (60.0, 723415), (61.0, 726032), (62.0, 728442), (63.0, 730629), (64.0, 732578), (65.0, 734280), (66.0, 735730), (67.0, 736928), (68.0, 737887), (69.0, 738627), (70.0, 739172), (71.0, 739540), (72.0, 739743), (73.0, 739792), (74.0, 739697), (75.0, 739472), (76.0, 739123), (77.0, 738658), (78.0, 738083), (79.0, 737406), (80.0, 737406)	(1.16)
World population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 3e+09), (1.00, 3.1e+09), (2.00, 3.1e+09), (3.00, 3.2e+09), (4.00, 3.3e+09), (5.00, 3.3e+09), (6.00, 3.4e+09), (7.00, 3.5e+09), (8.00, 3.6e+09), (9.00, 3.6e+09), (10.0, 3.7e+09), (11.0, 3.8e+09), (12.0, 3.9e+09), (13.0, 3.9e+09), (14.0, 4e+09), (15.0, 4.1e+09), (16.0, 4.2e+09), (17.0, 4.2e+09), (18.0, 4.3e+09), (19.0, 4.4e+09), (20.0, 4.5e+09), (21.0, 4.5e+09), (22.0, 4.6e+09), (23.0, 4.7e+09), (24.0, 4.8e+09), (25.0, 4.9e+09), (26.0, 4.9e+09), (27.0, 5e+09), (28.0, 5.1e+09), (29.0, 5.2e+09), (30.0, 5.3e+09), (31.0, 5.4e+09), (32.0, 5.4e+09), (33.0, 5.5e+09), (34.0, 5.6e+09), (35.0, 5.7e+09), (36.0, 5.8e+09), (37.0, 5.8e+09), (38.0, 5.9e+09), (39.0, 6e+09), (40.0, 6.1e+09), (41.0, 6.2e+09), (42.0, 6.2e+09), (43.0, 6.3e+09), (44.0, 6.4e+09), (45.0, 6.5e+09), (46.0, 6.5e+09), (47.0, 6.6e+09), (48.0, 6.7e+09), (49.0, 6.8e+09), (50.0, 6.8e+09), (51.0, 6.9e+09), (52.0, 7e+09), (53.0, 7e+09), (54.0, 7.1e+09), (55.0, 7.2e+09), (56.0, 7.2e+09), (57.0, 7.3e+09), (58.0, 7.4e+09), (59.0, 7.5e+09), (60.0, 7.5e+09), (61.0, 7.6e+09), (62.0, 7.6e+09), (63.0, 7.7e+09), (64.0, 7.8e+09), (65.0, 7.8e+09), (66.0, 7.9e+09), (67.0, 8e+09), (68.0, 8e+09), (69.0, 8.1e+09), (70.0, 8.1e+09), (71.0, 8.2e+09), (72.0, 8.3e+09), (73.0, 8.3e+09), (74.0, 8.4e+09), (75.0, 8.4e+09), (76.0, 8.5e+09), (77.0, 8.5e+09), (78.0, 8.6e+09), (79.0, 8.6e+09), (80.0, 8.7e+09)	(1.17)

stocks to change. The arrive rate of Fig. 4b, for example, shows tourists who flow into the tourism stock per month. Single-line arrows represent information flows, while plain text or circular icons depict auxiliary constant or converter variables, i. e., behavioral relations or decision points that convert information into decisions. Changes in the tourism stock, for example, depend on annual tourism, adjusted by tourism seasonality. Both the diagram on Fig. 4a and Table 1 are reproduced from the actual simulation model, first built on the glass of a computer screen using the diagramming interface of *iThink*® [67], and then specifying simple algebraic equations and parameter values. Built-in functions help quantify policy parameters and variables pertinent to Cyprus' tourism situation.

There is a one-to-one correspondence between the model diagram on Fig. 4a and its equations (Table 1). Like the diagram, the equations are the actual output from *iThink*® too. The equations corresponding to Fig. 8b are archived in [30]. Together, Cyprus' population, local tourism and monthly tourism determine the population and tourism sum (Eq. 1.11, Table 1). According to CYSTAT [11], both *Cyprus' Tourism Organization* and its government attach great importance to local tourism. A study on domestic tourism conducted in 1995 revealed that about 46 percent of Cypriots take long holidays. Of these, 61 percent take long holidays exclusively in Cyprus and eight percent in Cyprus and abroad, while 31 percent chose to travel abroad only. These are precisely the percentages in the model (Eq. 1.10, Table 1).

On Fig. 4a, the world land and population data, minus Cyprus' land, population and tourism co-determine the world EF (environmental footprint, Eq. 1.14, Table 1). Compared to Cyprus' smooth EF, i. e., the smooth ratio of the island's free land divided by its total population and tourism, the world EF gives a dynamic measure of Cyprus' relative attractiveness to the rest of the world. The EF ratio (Eq. 1.6, Table 1), i. e., the ratio of Cyprus' smooth EF (Eq. 1.13, Table 1) divided by the world EF (Eq. 1.14, Table 1), assumes that the higher this ratio is, the more attractive the island is to potential tourists, and vice versa. The EF ratio, which depends on Cyprus' total population and tourism, feeds back to the island's annual tourism via the inflow of foreign visitors who come to visit Cyprus every year (Fig. 4b).

The logistic or Verhulst growth model, after François Verhulst who published it in 1838 [66], helps explain Cyprus' actual annual tourism, a quantity that cannot grow forever (Fig. 4b). Every system that initially grows exponentially eventually approaches the carrying capacity of its environment, whether it is food supply for moose, the number of people susceptible to infection or the potential market for a good or a service. As an 'autopoietic' system approaches its limits to growth, it goes through a non-linear transition from a region where positive feedback dominates to a negative feedback dominated regime. S-shaped growth often results: a smooth transition from exponential growth to equilibrium.

The logistic model conforms to the requirements for S-shaped growth and the ecological idea of carrying capacity. The population it models typically grows in a fixed environment, such as Cyprus' foreign annual tourism has done since 1960 up to 2000. Initially dominated by positive feedback, Cyprus' annual tourism might soon reach the island's carrying capacity, with a nonlinear shift to dominance by negative feedback. While accounting for Cyprus' tourism lost to the summer of 1974 Turkish invasion, officially a very long 'military intervention', further depleting annual tourism is the outflow of Cyprus' visitors (not shown here) who might go as the island's free area reaches its Carrying Capacity, estimated at seventy times the number of Cyprus' visitors in 1960 [30].

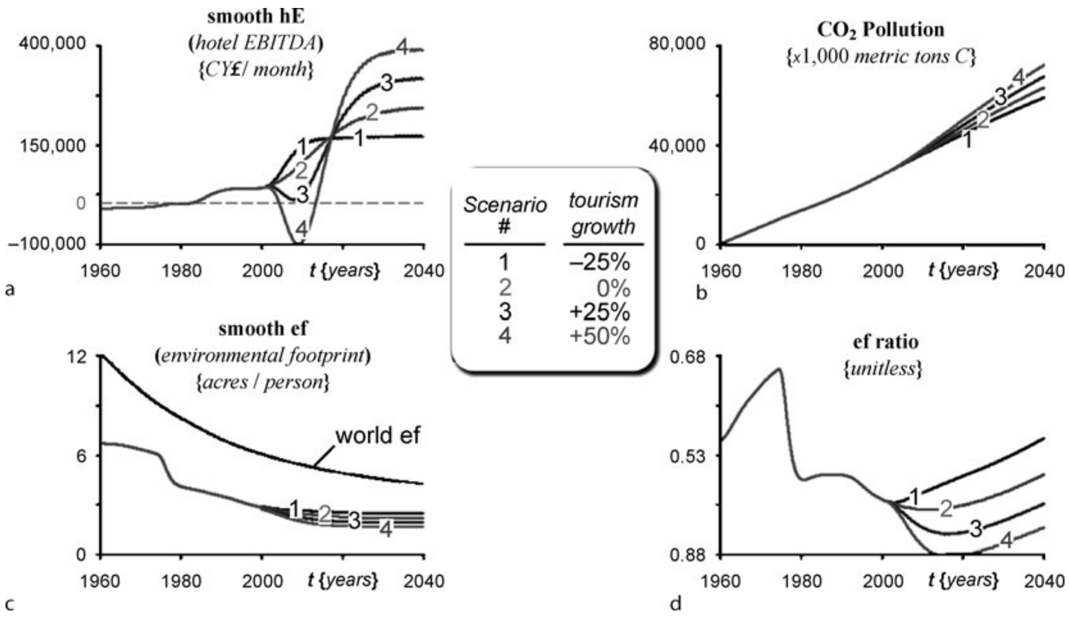
Cyprus' Hotel Association listed Cyprus tourism seasonality as one of its major concerns. At the time of this investigation, CYSTAT [11] had compiled monthly tourism data for only 30 months. These were used for computing Cyprus' tourism seasonality (Fig. 4b). Incorporating both the foreign annual tourism and the monthly tourism stocks in the model allows both looking at the big picture of annual tourism growth and assessing the potential long-term effects of tourism seasonality on the sustainability of Cyprus' environment and hotel EBITDA, i. e., *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization. The publicly available actual annual tourism data allow testing the model's usefulness, i. e., how faithfully it reproduces actual data between 1960 and 2000 [30].

Cyprus' foreign visitors and local tourists arrive at the island's hotels and resorts according to Cyprus' tourism seasonality, thereby feeding Cyprus' monthly tourism stock. About 11.3 days later, according to CYSTAT's [11] estimated average stay days, both foreign visitors and local tourists depart, thereby depleting the monthly tourism stock. By letting tourism growth = 0 and Cyprus' tourism seasonality continue repeating its established pattern, the model computes a zero-growth or *base-run* scenario. Subsequently, however, tourism growth values other than zero initiate different scenarios.

Cyprus' Tourism Growth Scenarios (Case 1)

What can Cyprus' hoteliers expect to see in terms of bottom-line dynamics? According to the four tourism-growth scenarios computed on Fig. 5, seasonal variations notwithstanding, the higher Cyprus' tourism growth is, the lower hotel EBITDA (smooth hE) is, in the short term. In the long term, however, higher tourism growth yields higher profitability in constant year 2000 prices.

High tourism growth implies accommodating overbooked hotel reservations for tourists who actually show up. Free cruises erode Cyprus' hotel EBITDA. The alternative is, however, angry tourists going off in hotel lobbies. Tourists have gotten angry at hotels before, but hotels have made the problem worse in recent years worldwide [16]. They have tightened check-in rules, doubled their reno-



Scenario-Driven Planning with System Dynamics, Figure 5

Four computed scenarios show how tourism growth might affect Cyprus' hotel EBITDA (smooth hE) and the island's environment, with carbon-dioxide (CO_2) pollution (adapted from and extending [30])

vations and increased the rate of overbooking by about 30 percent. The results can be explosive if one adds the record flight delays that travelers endure. Anyhow, free cruises to nearby Egypt and Israel sound much better than simply training employees to handle unhappy guests that scream in hotel lobbies.

Eventually, as Cyprus' bed capacity increases and thereby catches up with tourism demand, there will be less overbooking and a few free cruises to erode Cyprus' hotel EBITDA. Given enough time for an initial bed capacity disequilibrium adjustment, in the long term, high tourism growth increases both hotel EBITDA (Fig. 5a) and cash [30].

In addition to their profound consequences for its hotel value-chain participants, Cyprus' tourism growth might also determine the fate of the island's environment. Depending on the island's population and emissions per person, high tourism growth implies high anthropogenic emissions feeding Cyprus' CO_2 Pollution. Anthropogenic CO_2 emissions attributed to the upward and downward movements of recurring tourist arrivals create much more stress and strain for the island's natural environment than a consistent stream of tourism with low seasonality would. High tourism growth lowers Cyprus' environmental footprint (EF). The summer 1974 Turkish *military intervention* has had a drastic negative effect on Cyprus' relative attractiveness because it reduced the island's free land by 41 percent.

Although qualitatively similar to the world's average EF after the invasion, Cyprus' environmental footprint is lower than the world's average EF (Fig. 5c), rendering the island's free area relatively less attractive as more foreign tourists visit. Manifested in the EF ratio (Fig. 5d), Cyprus becomes relatively less attractive as more visitors choose to vacation on the island's free area.

Qualitatively, Cyprus' CO_2 pollution scenarios (Fig. 5b) look exactly like the A2 scenario family of harmonized anthropogenic CO_2 emissions, which the *Intergovernmental Panel on Climate Change* (IPCC) computed to access the risks of human-induced climate change [54]. Like in the rest of the world, unless drastic changes in policy or technology alter the emissions per person ratio in the next 40 years, CO_2 pollution is expected to grow proportionally with Cyprus' tourism, degrading the island's environment.

Case 2: A Japanese Chemicals Keiretsu (JCK)

Home of NASA's *Johnson Space Center*, the Clear Lake region in Texas boasts strong high technology, biotechnology and specialty chemicals firms. Among them is JCK, whose recent investment helps the Clear Lake region continue its stalwart role in Houston's regional economic expansion [40].

An active member of a famous Japanese giant conglomerate, JCK's history begun in the late 1800s. Despite its long history, however, it has not been easy for JCK to evade the feedback loop that drives Japanese firms to manufacture outside Japan. Since the 1950s, with Japan still recovering from WWII, the better Japanese companies performed, the better their national currency did. But the better Japan's currency did, the harder it became for its firms to export. The higher the yen, the more expensive and, therefore, less competitive Japan's exports become. This simple loop explains JCK's manufacturing lineage from Japan to USA [34].

But the transition process behind this lineage is not that simple. JCK's use of SdP with SD reveals a lot about its strategy design and implementation tactics. The model below shows a tiny fragment of JCK's gigantic effort to reperceive itself. The firm wants to see its keiretsu transform into an agile, virtual enterprise network (VEN) of active agents that collaborate to achieve its transnational business goals. Although still flying low under the media's collective radar screen, VENs receive increased attention by strategic managers [29].

Sterman (see Chap. 17 and 18 in [75]) presents a generic value-chain management structure that can unearth what VENs are about. By becoming a VEN, JCK is poised to bring the necessary people and production processes together to form *autopoietic*, i. e., self-organizing, customer-centric value chains in the specialty chemicals industry. JCK decided to build its own plant in USA because the net present value (NPV) of the anticipated combined cash flow resulting from a merger with other specialty chemicals manufacturers in USA would have been less then the sum of the NPVs of the projected cash flows of the firms acting independently. Moreover, JCK's own technology transfer cost is so low that the internalization cost associated with a merger would far exceed supplier charges plus market transaction costs. To remain competitive [62], JCK will not integrate the activity but offshoot it as a branch of its VEN-becoming keiretsu. The plant will be fully operational in January 2008. In order to maximize the combined *n*et *p*resent *v*alue of *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization, i. e., NPV(EBITDA), of its new USA plant and the existing one in Asia, JCK wishes to improve its USA sales revenue before production starts in USA.

JCK's pre-production marketing tactics aim at building a sales force to increase sales in USA. Until the completion of the new plant (Dec. 2007), JCK will keep importing chemicals from its plant in Asia. Once production starts in USA (Jan. 2008), then the flow of goods from Asia to USA stops, the plant in USA supplies the USA market and the flow of goods from USA to Asia begins.

Strategic scenarios are not new to the chemical industry [82]. SdP with SD helps this specialty chemicals producer integrate its business intelligence efforts with strategy design in anticipation of environmental change. Modeling JCK's strategic situation requires a comprehensive inquiry into the environmental causalities and equivocalities that dictate its actions. Computed strategic and tactical scenarios probe the combined consequences of environmental trends, changes in JCK's own strategy, as well as the moves of its current and future competitors. The section below describes briefly how JCK plans to implement its transnational strategy of balanced marketing and production. This takes the form of a system dynamics simulation model, which precedes the interpretation of its computed scenarios.

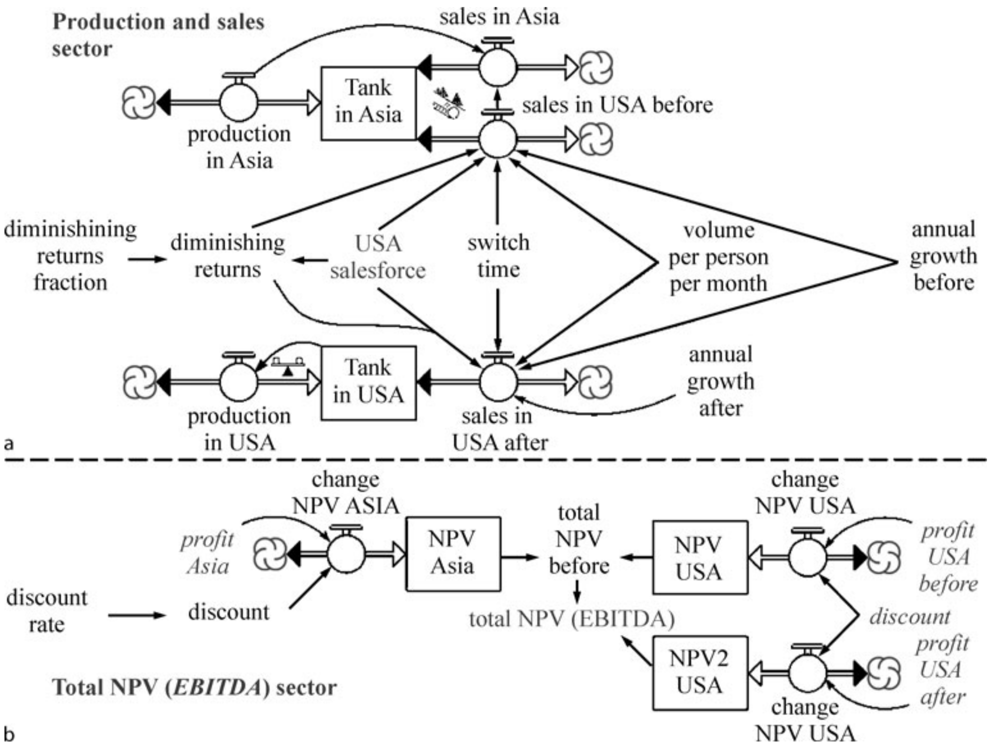
Model Description (Case 2)

The entire model has multiple sectors, four of which compute financial accounting data. Figure 6a shows the production and sales, and Fig. 6b the total NPV(EBITDA) model sectors. The corresponding algebra is in [34]. While JCK is building its USA factory, its factory in Asia makes and sells all specialty chemicals the USA market cannot yet absorb. This is what the *feed-forward* link from the production in Asia flow to the sales in Asia rate shows. The surplus demand JCK faces in Asia for its fine chemicals accounts for this rather unorthodox model structure. The surplus demand in Asia is the model's enabling *safety valve*, i. e., a major strategic assumption that renders tactical implementation feasible.

With the plant in Asia producing at full capacity before the switch, sales in the USA both depletes the tank in Asian stock and reduces sales in Asia. USA sales depend on JCK's USA sales force. But the size of this decision variable is just one determinant of sales in USA.

Sales productivity depends on many parameters, such as the annual growth before the switch rate of specialty chemicals in USA, the average expected volume a salesperson can sell per month as well as on the diminishing returns that sales people experience after the successful calls they initially make on their industrial customers. B2B or business to business, i. e., industrial marketing, can sometimes be as tough as B2C or business to customer, i. e., selling retail.

Time $t = 30$ months corresponds to January 2008, when the switch time converter cuts off the supply of JCK's chemicals from its plant in Asia. Ready by December 2007, the factory in the USA can supply the entire customer base its USA sales force will have been building for 30 months. As production in the USA begins, the



Scenario-Driven Planning with System Dynamics, Figure 6
JCK's a production and sales, and b total NPV(EBITDA) model sectors (adapted from [34]; NPV = net present value, and EBITDA = earnings before interest, taxes, depreciation and amortization)

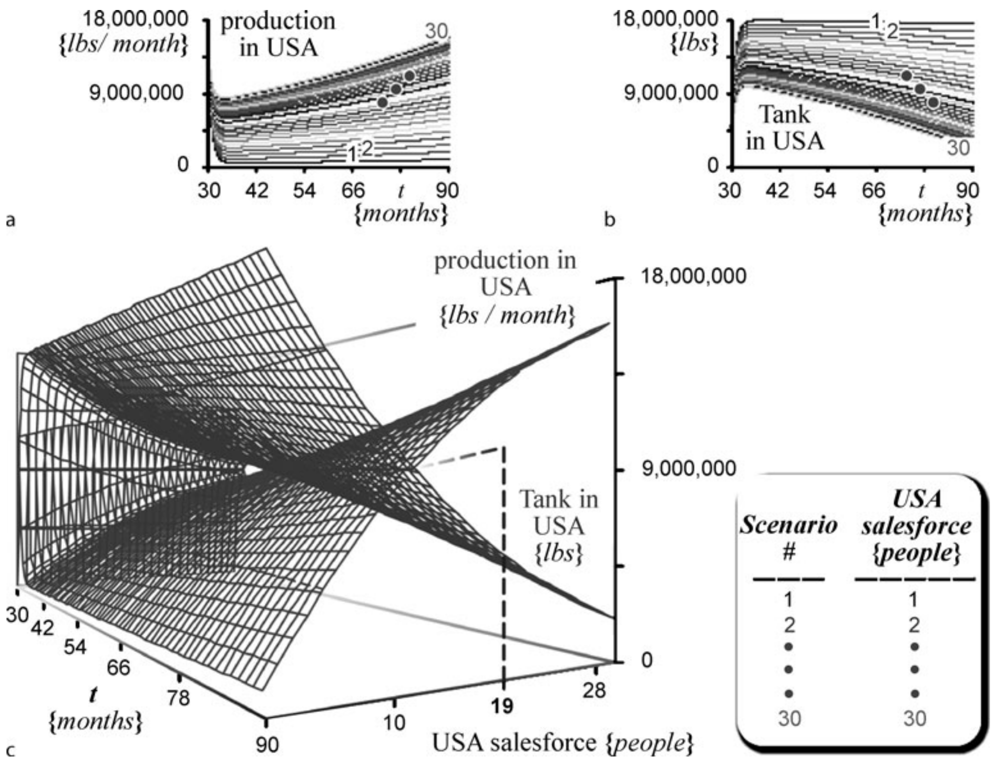
sales in the USA before flow stops draining the tank in Asia and sales in Asia resume to match JCK's surplus demand there. Acting both as a production flow and as a continuous-review inventory order point, after January 2008, production in USA feeds the tank in USA stock of the rudimentary value-chain management structure on Fig. 6a.

Value chains entail stock and flow structures for the acquisition, storage and conversion of inputs into outputs, and the decision rules that govern the flows. The jet ski value chain includes, for example, hulls and bows that travel down monorail assembly paths. At each stage in the process, a stock of parts buffers production. This includes the inventory of fiberglass laminate between hull and bow acquisition and usage, the inventory of hulls and bows for the jet ski lower and upper structures, and the inventory of jet skis between dealer acquisition and sales. The decision rules governing the flows entail policies for ordering fiberglass laminate from suppliers, scheduling the spraying of preformed molds with layers of fiberglass laminate before assembly, shipping new jet skis to dealers and customer demand.

A typical firm's or VEN's value chain consists of cascades of supply chains, which often extend beyond a firm's boundaries. Effective value chain models must incorporate different agents and firms, including suppliers, the firm, distribution channels and customers. Scenario-driven planning with system dynamics is well suited for value chain modeling and policy design because value chains involve multiple stock and flow chains, with time lags or delays, and the decision rules governing the flows create multiple feedback loops among VEN members or value- and supply-chain partners (see Chap. 17 and 18 in [75]).

Back to JCK, its tank in the USA feeds information about its level back to production in the USA. Acting first as a decision point, production in the USA compares the tank in the USA level to the tank's capacity. If the tank is not full, then production in the USA places an order to itself and, once the USA factory has the requisite capacity, production in the USA refills the tank in the USA, but only until sales in the USA after the switch drains the tank. Then the cycle begins all over again.

Meanwhile, the profit in Asia, profit in the USA before and profit in the USA after sectors [34] perform all the fi-



Scenario-Driven Planning with System Dynamics, Figure 7
Thirty computed scenarios show JCK's dual, smooth-switch and profitable purpose in production (adapted from [34])

nancial accounting necessary to keep track of the transactions that take place in the value chain production and sales sector (Fig. 6a). As each scenario runs, the profit in Asia, the USA before and the USA after sectors feed the corresponding change in net present value (NPV) flows of the model's total NPV(EBITDA) sector (Fig. 6b). By adjusting each profit sector's EBITDA according to the discount rate, the change in NPV flows compute the total NPV(EBITDA) both in Asia and in the USA, both before and after JCK's January 2008 supply switch.

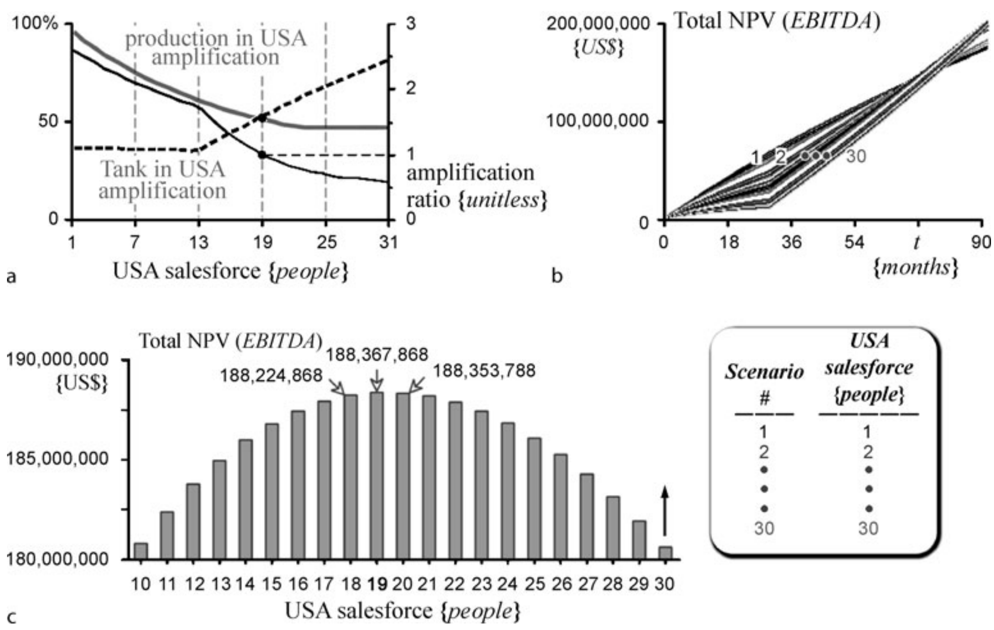
JCK's Computed Scenarios (Case 2)

Recall that the SdP with SD modeling-process spiral enabled our modeling team to crystallize JCK's strategic situation into the cyclical pattern that Fig. 3a shows. Although heavily disguised, the JCK measurement data and econometric sales functions let the system dynamics model compute scenarios to answer that razor-sharp optimization question the JCK executives asked:

> What size a USA sales force must we build in order to get a smooth switch in both sales and production in January 2008, and also to maximize the combined NPV(EBITDA) at our two plants in Asia and USA from now through 2012?

Treating the USA sales force policy parameter in the 'Sensi Specs…' menu item of *iThink*® allowed computing a set of 30 strategic scenarios. The 30 scenarios correspond to JCK's hiring from one to 30 sales people, respectively, to sell specialty chemicals to manufacturing firms in the USA, both before and after the January 2008 switch. Figures 7 and 8 show the 30 computed scenarios.

Figure 7c shows the response surfaces the production in USA rate and tank in the USA stock form after January 2008, in response to the 30 computed scenarios. The computed scenario that corresponds to JCK's building a USA sales force of 19 people achieves a smooth balance between sales in Asia and in the USA. Under this scenario, after January 2008, on the line where the two surfaces cross each other, not only the number of pounds of chemicals made and sold in Asia equals the number of pounds of chemicals made and sold in the USA, but as Fig. 7c shows, production in the USA also equals the tank in USA stock. So hiring 19 sales people now meets JCK's smooth switch in sales and production objective. But how?



Scenario-Driven Planning with System Dynamics, Figure 8
Thirty computed scenarios show how hiring a sales force of 19 people in the USA might maximize JCK's NPV(EBITDA), and thereby fulfill its dual, smooth-switch and profitable purpose (adapted from [34])

How does producing and selling in the USA at rates equal to the corresponding rates in Asia constitute a fair response to JCK's smooth switch objective? The JCK executives seemed to accept this at face value. But our team had to clearly explain the dynamics of JCK's rudimentary USA value chain (Fig. 6a), in order to unearth what the USA member of this transnational VEN-becoming keiretsu might be up to.

It looks simple, but the value chain of the production and sales sector on Fig. 6a can show the same amplification symptoms that much more elaborate value chains show when they fall pray to bullwhip effects. Locally rational policies that create smooth and stable adjustment of individual business units can, through their interaction with other functions and firms, cause oscillation and instability. Figure 8a shows the profound consequences of JCK's switch for its value chain in the USA. Because of the sudden switch in January 2008, the computed scenarios cause 30 sudden step changes. Both variables' adjustment rates increase, but the tank in the USA stock's amplification remains almost constant below 50 percent. As customer demand steps up, so do both metrics' new equilibrium points, but in direct proportion to the step increase in customer demand in the USA.

The 30 computed scenarios confirm Sterman's argument that, while the magnitude of amplification depends on stock adjustment times and delivery lags, its existence does not. No matter how drastically customers and firms downstream in a value chain change an orders' magnitude, they cannot affect supply chain amplification. Value chain managers must never blame customers and downstream firms or their forecasts for bullwhip effects. The production in USA amplification is almost double the tank in USA for a small USA sales force, suggesting that JCK's USA factory faces much larger changes in demand than its sales people do. Although temporary, during its disequilibrium adjustment, the tank in the USA consistently overshoots the new equilibrium points that it seeks after the switch (Fig. 7b), an inevitable consequence of stock and flow structure. Customers are innocent, but JCK's value chain structure is not:

First, the tank in the USA stock adjustment process creates significant amplification of the production in the USA rate. Though the tank in the USA relative amplification is 36.18 percent under the USA sales force = 1 scenario, for example, the relative amplification of production in the USA (Fig. 8a) increases by a maximum of more than 90 percent (the peak production in the USA rate, after t = 30 months, divided by the minimum production in the USA rate $= 11,766,430.01/1,026,107.64 = 91.28$ percent). The *amplification ratio*, i. e., the ratio of maximum change in output to maximum change in input, therefore is $91.28\%/36.18\% = 2.52$. A one-percent increase in demand for JCK's chemicals causes a 2.52 percent surge in demand at JCK's USA plant. While the amplification ratio magnitude depends on the stock adjustment times and delivery lags, its existence does not (see p. 673 in [75]).

Second, amplification is temporary. In the long run, a one-percent increase in sales in the USA after leads to a one-percent increase in production in the USA. After two-adjustment times, i. e., two months, production in the USA gradually drops (Fig. 7a). During the disequilibrium adjustment, however, production in the USA overshoots its new equilibrium, an inevitable consequence of the stock and flow structure of customer-supplier value chains, no matter how tiny or simple they are. The only way the tank in the USA stock can increase is for its inflow production in the USA rate (order rate) to exceed its outflow rate sales in the USA after (Fig. 6a). Within a VEN's or keiretsu's customer-supplier value chain, supply agents face much larger changes in demand than finished-goods inventory, and the surge in demand is temporary.

The computed scenarios show that as the USA sales force increases, the production in the USA's rate of amplification declines because its new long-term equilibrium point is closer to its initial jump in January 2008. Conversely, as the tank in the USA stock's long-term equilibrium point remains consistently high because of the larger USA sales force, its relative amplification begins to rise. Since the two variables' relative amplification moves in opposite directions, eventually, they meet. What a coincidence! They meet above the USA sales force = 19 people. Now, is this not a much better interpretation of the word 'smooth' in fair response to JCK's smooth-switch performance purpose? The answer to JCK is now pertinent to its balancing its value chain in USA. With a USA sales force = 19, JCK's value chain components show equal relative amplification to sudden changes in demand, attaining nothing less than a magnificent amplification ratio = 1. Now that is smooth!

But what of profitability? JCK's polite executives said: "maximize… combined… NPV". In the time domain (Fig. 8b), total NPV(EBITDA) creates intricate dynamics that obscures the USA sales force effect. But the phase plot on Fig. 8c clearly shows a concave down behavior along the USA sales force: USA sales force = 19 maximizes the two plants' combined total NPV(EBITDA).

Future Directions

The above cases show how scenario-driven planning with system dynamics helps control performance by enabling organizational learning and the management of uncertainty. The strategic intelligence system that SdP with SD provides rests on the idea of a collective inquiry, which translates the environmental 'macrocosm' and a firm's 'microcosm' into a shared causal map with computed scenarios. Informed discussion then takes place. Seeing SdP with SD as an inquiry system might help the outcomes of the situation formulation-solution-implementation sequence, each stage built on successive learning.

Strategic situations are complex and uncertain. Because planning is directed toward the future, predictions of changes in the environment are indispensable components of it. Conventional forecasting by itself provides no cohesive way of understanding the effect of changes that might occur in the future. Conversely, SdP with SD and its computed scenarios provide strategic intelligence and a link from traditional forecasting to modern interactive planning systems. In today's quest for managers who are more leaders than conciliators, the strategists' or executives' interest in scenarios must be welcomed. A clearer delineation of SdP with SD might make it a very rich field of application and research.

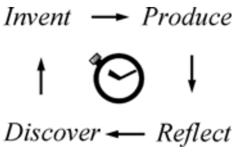
The SdP with SD inquiry system on Fig. 2 includes several contributions. *First*, by translating the environmental *macrocosm* and the firm *microcosm* into a common context for conceptualization, the requisites of theory building can be addressed. Planning analysts no longer have to operate piecemeal. A theory and a dominant logic typically emerge from shared perceptions about a firm, its environment and stakeholder purposes through model construction.

Second, the outputs of the strategic management process activities build on each other as successive layers. The SAST loop on Fig. 2 follows the counterclockwise direction of multiperspective dialectics [47]. This process allows adjustment of individual and organizational theories and logic, leading to an evolutionary interpretation of the real system that strategic decisions target.

Third, the inquiry system of Fig. 2 enables flexible support for all phases of strategy design. Problem finding or forming, or situation formulation receives equal attention as problem solving.

SdP with SD helps open up the black box of decision makers' mental models, so they can specify the ideas and rules they apply. That in turn helps enrich their language and label system, organizational capability and knowledge, and strategic decision processing system. Computed scenarios bring about transformation rules not previously thought of as well as new variables and interaction paths.

As an entity, each decision maker has a local scope and deals only with specific variables and access paths to other entities. But success factors are not etched in stone. Often, we only observe a representative state of each entity, namely, locally meaningful variables and parts of a sce-



nario. This representativeness changes dynamically in the process of computing scenarios. Beyond the purely technical advantages of computed scenarios, planning becomes interactive, and language and label systems render themselves more adequate, effective and precise. Their associated organizational capability develops even more. In addition, the minor and major assumptions in decision makers' mental models surface as computed scenarios specify the conditions under which performance changes.

A line of great immediate concern requires researchers and practitioners alike to explore the modeling process behind SdP with SD. For the sake of realism, to make negotiated perceptions of reality explicit, we need representations where strategic real options and self-interest projections mold the way in which managers incorporate their observations and interpretations into strategy models. This is an unavoidable, most challenging path to tread, if we want to build a dialectical debate into the strategy design process.

Do we really want to? Yes because:

1. The traditional hierarchical organization dogma has been planning, managing and controlling, whereas the new reality of the learning organization incorporates vision, values and mental models. It entails training managers and teams in the IPRD learning cycle conceived by Dewey [14] (cf. Senge and Sterman [71]):
2. In the strategic management process (SMP) evolution, planning is evolving too, from objective-driven to budget-driven to strategy-driven to scenario-driven planning with system dynamics (SdP with SD, see pp. 271–272 in [32]).
3. The inquiry system that mediates the restructuring of organizational *theory in use* [68] determines the quality of organizational learning.

By looking into the dynamics of strategy design and the resulting performance of firms, the SdP with SD framework on Fig. 2 might let managers, planners and business researchers see the tremendous potential of computed strategic scenarios. They might choose to build intelligence systems around SdP with SD to create insight for strategy design. They will be building real knowledge in the process, while developing capability for institutional learning. Both Pascale [59] and de Geus [13] see the capability to speed up institutional learning as a truly sustainable competitive advantage.

Bibliography

Primary Literature

1. Acar W (1983) Toward a theory of problem formulation and the planning of change: Causal mapping and dialectical debate in situation formulation. UMI, Ann Arbor
2. Ackoff RL (1981) Creating the corporate future. Wiley, New York
3. Ackoff RL, Emery FE (1972) On purposeful systems. Aldine-Atherton, Chicago
4. Amara R, Lipinski AJ (1983) Business planning for an uncertain future: Scenarios and strategies. Pergamon, New York
5. Anderson TJ (2000) Real options analysis in strategic decision making: An applied approach in a dual options framework. J Appl Manag Stud 9(2):235–255
6. Ansoff HI (1985) Conceptual underpinnings of systematic strategic management. Eur J Oper Res 19(1):2–19
7. Ansoff HI, McDonnell E (1990) Implanting strategic management, 2nd edn. Prentice-Hall, New York
8. Brauers J, Weber M (1988) A new method of scenario analysis for strategic planning. J Forecast 7(1):31–47
9. Brenkert AL (1998) Carbon dioxide emission estimates from fossil-fuel burning, hydraulic cement production, and gas flaring for 1995 on a one degree grid cell basis. Oak Ridge National Laboratory, Carbon Dioxide Information Analysis Center, Oak Ridge, TN (Database: NDP-058A 2-1998)
10. Christensen CM (1997) The innovator's dilemma: When new technologies cause great firms to fail. Harvard Business School Press, Cambridge
11. CYSTAT (2000) Cyprus key figures: Tourism. The Statistical Service of Cyprus (CYSTAT), Nicosia
12. Daft RL, Weick KE (1984) Toward a model of organizations as interpretation systems. Acad Manag Rev 9:284–295
13. de Geus AP (1992) Modelling to predict or to learn? Eur J Oper Res 59(1):1–5
14. Dewey J (1938) Logic: The theory of inquiry. Holt, Rinehart and Winston, New York
15. Donaldson L (1992) The Weick stuff: Managing beyond games. Organ Sci 3(4):461–466
16. Drucker J, Higgins M (2001) Hotel rage: Losing it in the lobby. Wall Street J (Fri 16 Feb):W1–W7
17. Duncan RB (1972) Characteristics of organizational environments and perceived environmental uncertainty. Adm Sci Q 17:313–327
18. Eberlein RL (2002) Vensim® PLE Software, V 5.2a. Ventana Systems Inc, Harvard
19. Eden C (1994) Cognitive mapping and problem structuring for system dynamics model building. Syst Dyn Rev 10(3):257–276
20. Eden C (2004) Analyzing cognitive maps to help structure issues or problems. Eur J Oper Res 159:673–686
21. Eilon S (1984) The Art of Reckoning: Analysis of performance criteria. Academic Press, London
22. Emery FE, Trist EL (1965) The causal texture of organizational environments. Hum Relat 18:21–32
23. Forrester JW (1958) Industrial dynamics: A major breakthrough for decision makers. Harvard Bus Rev 36(4):37–66

24. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
25. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3(2):136–149
26. Forrester JW (1992) Policies, decisions and information sources for modeling. Eur J Oper Res 59(1):42–63
27. Georgantzas NC (1995) Strategy design tradeoffs-free. Hum Syst Manag 14(2):149–161
28. Georgantzas NC (2001) Simulation modeling. In: Warner M (ed) International encyclopedia of business and management, 2nd edn. Thomson Learning, London, pp 5861–5872
29. Georgantzas NC (2001) Virtual enterprise networks: The fifth element of corporate governance. Hum Syst Manag 20(3):171–188, with a 2003 reprint ICFAI J Corp Gov 2(4):67–91
30. Georgantzas NC (2003) Tourism dynamics: Cyprus' hotel value chain and profitability. Syst Dyn Rev 19(3):175–212
31. Georgantzas NC (2007) Digest® wisdom: Collaborate for win-win human systems. In: Shi Y et al (eds) Advances in multiple criteria decision making and human systems management. IOS Press, Amsterdam, pp 341–371
32. Georgantzas NC, Acar W (1995) Scenario-driven planning: Learning to manage strategic uncertainty. Greenwood-Quorum, Westport
33. Georgantzas NC, Katsamakas E (2007) Disruptive innovation strategy effects on hard-disk maker population: A system dynamics study. Inf Resour Manag J 20(2):90–107
34. Georgantzas NC, Sasai K, Schrömbgens P, Richtenburg K et al (2002) A chemical firm's penetration strategy, balance and profitability. In: Proc of the 20th international system dynamics society conference, 28 Jul–1 Aug, Villa Igiea, Palermo, Italy
35. Gharajedaghi J (1999) Systems thinking: Managing chaos and complexity – A platform for designing business architecture. Butterworth-Heinemann, Boston
36. Godet M (1987) Scenarios and strategic management: Prospective et planification stratégique. Butterworths, London
37. Godet M, Roubelat F (1996) Creating the future: The use and misuse of scenarios. Long Range Plan 29(2):164–171
38. Hax AC, Majluf NS (1996) The strategy concept and process: A pragmatic approach, 2nd edn. Prentice Hall, Upper Saddle River
39. Helmer O (1983) Looking forward. Sage, Beverly Hills
40. Hodgin RF (2001) Clear Lake Area Industry and Projections 2001. Center for Economic Development and Research, University of Houston-Clear Lake, Houston
41. Huss WR, Honton EJ (1987) Scenario planning: What style should you use? Long Range Plan 20(4):21–29
42. Istvan RL (1992) A new productivity paradigm for competitive advantage. Strateg Manag J 13(7):525–537
43. Jarillo JC (1988) On strategic networks. Strateg Manag J 9(1):31–41
44. Jarillo JC, Martínez JI (1990) Different roles for subsidiaries: The case of multinational corporations in Spain. Manag J 11(7):501–512
45. Kahn H, Wiener AJ (1967) The next thirty-three years: A framework for speculation. Daedalus 96(3):705–732
46. Lissack MR, Roos J (1999) The next common sense: The e-manager's guide to mastering complexity. Nicholas Brealey Publishing, London
47. Mason RO, Mitroff II (1981) Challenging strategic planning assumptions. Wiley, New York
48. Miller D, Friesen PH (1983) Strategy making and environment: The third link. Strateg Manag J 4:221–235
49. Millet SM, Randles F (1986) Scenarios for strategic business planning: A case history for aerospace and defence companies. Interfaces. 16(6):64–72
50. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest® to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
51. Moore GA (1991) Crossing the chasm. Harper-Collins, New York
52. Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31:900–916
53. Morecroft JDW (1988) System dynamics and microworlds for policymakers. Eur J Oper Res 35:310–320
54. Nakicenovic N, Davidson O, Davis G et al (2000) Summary for policymakers–Emissions scenarios: A special report of working group III of the intergovernmental panel on climate change (IPCC). World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP), New York
55. Nicolis JS (1986) Dynamics of hierarchical systems: An evolutionary approach. Springer, Berlin
56. Oliva R (2004) Model structure analysis through graph theory: Partition heuristics and feedback structure decomposition. Syst Dyn Rev 20(4):313–336
57. Oliva R, Mojtahedzadeh MT (2004) Keep it simple: A dominance assessment of short feedback loops. In: Proc of the 22nd international system dynamics society conference, 25–29 July 2004, Keble College, Oxford University, Oxford UK
58. Ozbekhan H (1977) The future of Paris: A systems study in strategic urban planning. Philos Trans Royal Soc London. 387:523–544
59. Pascale RT (1984) Perspectives on strategy: The real story behind Honda's success. California Manag Rev 26(Spring):47–72
60. Pine BJ-I, Victor B, Boynton AC (1993) Making mass customization work. Harvard Bus Rev 71(5):108–115
61. Porter ME (1985) Competitive advantage: Creating and sustaining superior performance. Free Press, New York
62. Porter ME (1991) Towards a dynamic theory of strategy. Strateg Manag J 12(Winter Special Issue):95–117
63. Randers J (1980) Guidelines for model conceptualization. In: Randers J (ed) Elements of the system dynamics method. MIT Press, Cambridge, pp 117–139
64. Raynor ME (2007) The strategy paradox: Why committing to success leads to failure [And What to Do About It]. Currency-Doubleday, New York
65. Repenning NP (2003) Selling system dynamics to (other) social scientists. Syst Dyn Rev 19(4):303–327
66. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
67. Richmond B et al (2006) iThink® Software V 9.0.2. iSee Systems™, Lebanon
68. Schön D (1983) Organizational learning. In: Morgan G (ed) Beyond method. Sage, London
69. Schrage M (1991) Spreadsheets: Bulking up on data. San Francisco Examiner
70. Schwenk CR (1984) Cognitive simplification processes in strategic decision making. Strateg Manag J 5(2):111–128
71. Senge PM, Sterman JD (1992) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59(1):137–150
72. Simon HA (1979) Rational decision making in business organizations. Am Econ Rev 69(4):497–509
73. Singer AE (1992) Strategy as rationality. Hum Syst Manag 11(1):7–21

74. Singer AE (1994) Strategy as moral philosophy. Strateg Manag J 15:191–213
75. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
76. Sterman JD, Wittenberg J (1999) Path dependence, competition and succession in the dynamics of scientific revolution. Organ Sci 10(3, Special issue: Application of complexity theory to organization science, May–Jun):322–341
77. Turner F (1997) Foreword: Chaos and social science. In: Eve RA, Horsfall S, Lee ME (eds) Chaos, complexity and sociology. Sage, Thousand Oaks, pp xi–xxvii
78. Wack P (1985) Scenarios: Uncharted waters ahead. Harvard Bus Rev 63(5):73–89
79. Wack P (1985) Scenarios: Shooting the rapids. Harvard Bus Rev 63(6):139–150
80. Zeleny M (1988) Parallelism, integration, autocoordination and ambiguity in human support systems. In: Gupta MM, Yamakawa T (eds) Fuzzy logic in knowledge-based systems, decision and control. Elsevier Science, North Holland, pp 107–122
81. Zeleny M (2005) Human systems management: Integrating knowledge, management and systems. World Scientific, Hackensack
82. Zentner RD (1987) Scenarios and the chemical industry. Chem Marketing Manag (Spring):21–25

Books and Reviews

Bazerman MH, Watkins MD (2004) Predictable surprises: The disasters you should have seen coming and how to prevent them. Harvard Business School Press, Boston

Bower G, Morrow D (1990) Mental models in narrative comprehension. Science 247(4938):44–48

Mittelstaedt RE (2005) Will your next mistake be fatal? Avoiding the chain of mistakes that can destroy your organization. Wharton School Publishing, Upper Saddle River

Morecroft JDW (2007) Strategic modeling and business dynamics: A feedback systems approach. Wiley, West Sussex

Schnaars SP (1989) Megamistakes: Forecasting and the myth of rapid technological change. The Free Press, New York

Schwartz P (1991) The art of the long view. Doubleday-Currency, New York

Tuchman B (1985) The March of folly: From troy to vietnam. Ballantine Books, New York

Vennix JAM (1996) Group model building: Facilitating team learning using system dynamics. Wiley, Chichester

Stochastic Volatility

Torben G. Andersen[1,2,3] , Luca Benzoni[4]
[1] Kellogg School of Management, Northwestern University, Evanston, USA
[2] NBER, Cambridge, USA
[3] CREATES, Aarhus, Denmark
[4] Federal Reserve Bank of Chicago, Chicago, USA

Article Outline

Glossary

Implied volatility The value of asset return volatility which equates a model-implied derivative price to the observed market price. Most notably, the term is used to identify the volatility implied by the Black and Scholes [63] option pricing formula.

Quadratic return variation The ex-post sample-path return variation over a fixed time interval.

Realized volatility The sum of finely sampled squared asset return realizations over a fixed time interval. It is an estimate of the quadratic return variation over such time interval.

Stochastic volatility A process in which the return variation dynamics include an unobservable shock which cannot be predicted using current available information.

Definition of the Subject

Given the importance of return volatility on a number of practical financial management decisions, the efforts to provide good real-time estimates and forecasts of current and future volatility have been extensive. The main framework used in this context involves stochastic volatility models. In a broad sense, this model class includes GARCH, but we focus on a narrower set of specifications in which volatility follows its own random process, as is common in models originating within financial economics. The distinguishing feature of these specifications is that volatility, being inherently unobservable and subject to independent random shocks, is not measurable with respect to observable information. In what follows, we refer to these models as *genuine* stochastic volatility models.

Much modern asset pricing theory is built on continuous-time models. The natural concept of volatility within this setting is that of genuine stochastic volatility. For example, stochastic volatility (jump-)diffusions have provided a useful tool for a wide range of applications, including the pricing of options and other derivatives, the modeling of the term structure of risk-free interest rates, and the pricing of foreign currencies and defaultable bonds. The increased use of intraday transaction data for construction of so-called realized volatility measures provides additional impetus for considering genuine stochastic volatility models. As we demonstrate below, the realized volatility approach is closely associated with the continuous-time stochastic volatility framework of financial economics.

There are some unique challenges in dealing with genuine stochastic volatility models. For example, volatility is truly latent and this feature complicates estimation and inference. Further, the presence of an additional state variable – volatility – renders the model less tractable from an analytic perspective. We review how such challenges have been addressed through development of new estimation methods and imposition of model restrictions allowing for closed-form solutions while remaining consistent with the dominant empirical features of the data.

Introduction

The label Stochastic Volatility is applied in two distinct ways in the literature. For one, it is used to signify that the (absolute) size of the innovations of a time series displays random fluctuations over time. Descriptive models of financial time series almost invariably embed this feature nowadays as asset return series tend to display alternating quiet and turbulent periods of varying length and intensity. To distinguish this feature from models that operate with an a priori known or deterministic path for the volatility process, the random evolution of the conditional return variance is termed stochastic volatility. The simplest case of deterministic volatility is the constant variance assumption invoked in, e. g., the Black and Scholes [63] framework. Another example is modeling the variance purely as a given function of calendar time, allowing only for effects such as time-of-year (seasonals), day-of-week (institutional and announcement driven) or time-of-day (diurnal effects due to, e. g., market microstructure features). Any model not falling within this class is then a stochastic volatility model. For example, in the one-factor continuous-time Cox, Ingersoll, and Ross [113] (CIR) model the (stochastic) level of the short term interest rate governs the dynamics of the (instantaneous) drift and diffusion term of all zero-coupon yields. Likewise, in GARCH models the past return innovations govern the one-period ahead conditional mean and variance. In both models, the volatility is known, or deterministic, at a given point in time, but the random evolution of the processes renders volatility stochastic for any horizon beyond the present period.

The second notion of stochastic volatility, which we adopt henceforth, refers to models in which the return variation dynamics is subject to an unobserved random shock so that the volatility is inherently latent. That is, the current volatility state is not known for sure, conditional on the true data generating process and the past history of all available discretely sampled data. Since the CIR and GARCH models described above render the current (conditional) volatility known, they are not stochastic volatility models in this sense. In order to make the distinction clear cut, we follow Andersen [10] and label this second, more restrictive, set *genuine* stochastic volatility (SV) models.

There are two main advantages to focusing on SV models. First, much asset pricing theory is built on continuous-time models. Within this class, SV models tend to fit more naturally with a wide array of applications, including the pricing of currencies, options, and other derivatives, as well as the modeling of the term structure of interest rates. Second, the increasing use of high-frequency intraday data for construction of so-called realized volatility measures is also starting to push the GARCH models out of the limelight as the realized volatility approach is naturally linked to the continuous-time SV framework of financial economics.

One drawback is that volatility is not measurable with respect to observable (past) information in the SV setting. As such, an estimate of the current volatility state must be filtered out from a noisy environment and the estimate will change as future observations become available. Hence, in-sample estimation typically involves smoothing techniques, not just filtering. In contrast, the conditional variance in GARCH is observable given past information, which renders (quasi-)maximum likelihood techniques for inference quite straightforward while smoothing techniques have no role. As such, GARCH models are easier to estimate and practitioners often rely on them for time-series forecasts of volatility. However, the development of powerful method of simulated moments, Markov Chain Monte Carlo (MCMC) and other simulation based procedures for estimation and forecasting of SV models may well render them competitive with ARCH over time on that dimension.

Direct indications of the relations between SV and GARCH models are evident in the sequence of papers by Dan Nelson and Dean Foster exploring the SV diffusion limits of ARCH models as the case of continuous sampling is approached, see, e. g., Nelson and Foster [219]. Moreover, as explained in further detail in the estimation section below, it can be useful to summarize the dynamic features of asset returns by tractable pseudo-likelihood scores obtained from GARCH-style models when performing simulation based inference for SV models. As such, the SV and GARCH frameworks are closely related and should be viewed as complements. Despite these connections we focus, for the sake of brevity, almost exclusively on SV models and refer the interested reader to the GARCH chapter for further information.

The literature on SV models is vast and rapidly growing, and excellent surveys are available, e. g., Ghysels et al. [158] and Shephard [239,240]. Consequently, we focus on providing an overview of the main approaches with illustrations of the scope for applications of these models to practical finance problems.

Model Specification

The original econometric studies of SV models were invariably cast in discrete time and they were quite similar in structure to ARCH models, although endowed with a more explicit structural interpretation. Recent work in the area has been mostly directly towards a continuous time setting and motivated by the typical specifications in financial economics. This section briefly reviews the two alternative approaches to specification of SV models.

Discrete-Time SV Models and the Mixture-of-Distributions Hypothesis

Asset pricing theory contends that financial asset prices reflect the discounted value of future expected cash flows, implying that all news relevant for either discount rates or cash flows should induce a shift in market prices. Since economic news items appear almost continuously in real time, this perspective rationalizes the ever-changing nature of prices observed in financial markets. The process linking news arrivals to price changes may be complex, but if it is stationary in the statistical sense it will nonetheless produce a robust theoretical association between news arrivals, market activity and return volatility. In fact, if the number of news arrival is very large, standard central limit theory will tend to imply that asset returns are approximately normally distributed *conditional* on the news count. More generally, variables such as the trading volume, the number of transactions or the number of price quotes are also naturally related to the intensity of the information flow. This line of reasoning has motivated specifications such as

$$y_t|s_t \rightsquigarrow N(\mu_y s_t \ , \ \sigma_y^2 s_t) \ , \qquad (1)$$

where y_t is an "activity" variable related to the information flow, s_t is a positive intensity process reflecting the rate of news arrivals, μ_y represents the mean response to an information event, and σ_y is a pure scaling parameter.

This is a normal mixture model, where the s_t process governs or "mixes" the scale of the distribution across the periods. If s_t is constant, this is simply an i.i.d. Gaussian process for returns and possible other related variables. However, this is clearly at odds with the empirical evidence for, e. g., return volatility and trading volume. Therefore, s_t is typically stipulated to follow a separate stochastic process with random innovations. Hence, each period the return series is subject to two separate shocks, namely the usual idiosyncratic error term associated with the (normal) return distribution, but also a shock to the variance or volatility process, s_t. This endows the return process with genuine stochastic volatility, reflecting the random intensity of news arrivals. Moreover, it is typically assumed that only returns, transactions and quotes are observable, but not the actual value of s_t itself, implying that σ_y cannot be separately identified. Hence, we simply fix this parameter at unity.

The time variation in the information flow series induces a fat-tailed unconditional distribution, consistent with stylized facts for financial return and, e. g., trading volume series. Intuitively, days with a lot of news display more rapid price fluctuations and trading activity than days with a low news count. In addition, if the s_t process is positively correlated, then shocks to the conditional mean and variance processes for y_t will be persistent. This is consistent with the observed clustering in financial markets, where return volatility and trading activity are contemporaneously correlated and each display pronounced positive serial dependence.

The inherent randomness and unobserved nature of the news arrival process, even during period t, renders the true mean and variance series latent. This property is the major difference with the GARCH model class, in which the one-step-ahead conditional mean and variance are a known function of observed variables at time $t-1$. As such, for genuine SV models, we must distinguish the full, but infeasible, information set ($s_t \in \mathcal{F}_t$) and the observable information set ($s_t \notin \mathcal{I}_t$). This basic latency of the mixing variable (state vector) of the SV model complicates inference and forecasting procedures as discussed below.

For short horizon returns, μ_y is nearly negligible and can reasonably be ignored or simply fixed at a small constant value, and the series can then be demeaned. This simplification produces the following return (innovation) model,

$$r_t = \sqrt{s_t}\, z_t \ , \qquad (2)$$

where z_t is an i.i.d. standard normal variable, implying a simple normal-mixture representation,

$$r_t|s_t \rightsquigarrow N(0, s_t) \ . \qquad (3)$$

Univariate return models of the form (3) as well as multivariate systems including a return variable along with other related market activity variables, such as the transactions count, the quote intensity or the aggregate trading volume, stem from the Mixture-of-Distributions Hypothesis (MDH).

Actual implementation of the MDH hinges on a particular representation of the information-arrival process s_t. Clark [102] uses trading volume as a proxy for the activity variable, a choice motivated by the high contemporaneous correlation between return volatility and volume. Tauchen and Pitts [247] follow a structural approach to characterize the joint distribution of the daily return and volume relation governed by the underlying latent information flow s_t. However, both these models assume temporal independence of the information flow, thus failing to capture the clustering in these series. Partly in response, Gallant et al. [153] examine the joint conditional return-volume distribution without imposing any structural MDH restrictions. Nonetheless, many of the original discrete-time SV specifications are compatible with the MDH framework, including Taylor [249][1], who proposes an autoregressive parametrization of the latent log-volatility (or information flow) variable

$$\log(s_{t+1}) = \eta_0 + \eta_1 \log(s_t) + u_t, \ u_t \rightsquigarrow \text{i.i.d}(0, \sigma_u^2) \ , \qquad (4)$$

where the error term, u_t, may be correlated with the disturbance term, z_t, in the return Eq. (2) so that $\rho = \operatorname{corr}(u_t, z_t) \neq 0$. If $\rho < 0$, downward movements in asset prices result in higher future volatility as also predicted by the so-called 'leverage effect' in the exponential GARCH, or EGARCH, form of Nelson [218] and the asymmetric GARCH model of Glosten et al. [160].

[1] Discrete-time SV models go father back in time, at least to the easly paper by Rosenberg [232] recently reprinted in Shephard [240].

Early tests of the MDH include Lamoureux and Lastrapes [194] and Richardson and Smith [231]. Subsequently, Andersen [11] studies a modified version of the MDH that provides a much improved fit to the data. Further refinements of the MDH specification have been pursued by, e. g., Liesenfeld [198,199] and Bollerslev and Jubinsky [67]. Among the first empirical studies of the related approach of stochastic time changes are Ané and Geman [29], who focus on stock returns, and Conley et al. [109], who focus on the short-term risk-free interest rate.

Continuous-Time Stochastic Volatility Models

Asset returns typically contain a predictable component, which compensates the investor for the risk of holding the security, and an unobservable shock term, which cannot be predicted using current available information. The conditional asset return variance pertains to the variability of the unobservable shock term. As such, over a non-infinitesimal horizon it is necessary to first specify the conditional mean return (e. g., through an asset pricing model) in order to identify the conditional return variation. In contrast, over an infinitesimal time interval this is not necessary because the requirement that market prices do not admit arbitrage opportunities implies that return innovations are an order of magnitude larger than the mean return. This result has important implications for the approach we use to model and measure volatility in continuous time.

Consider an asset with log-price process $\{p(t)\ ,\ t \in [0, T]\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$. Following Andersen et al. [19] we define the continuously compounded asset return over a time interval from $t-h$ to t, $0 \le h \le t \le T$, to be

$$r(t, h) = p(t) - p(t-h)\ . \tag{5}$$

A special case of (5) is the cumulative return up to time t, which we denote $r(t) \equiv r(t,t) = p(t) - p(0)$, $0 \le t \le T$. Assume the asset trades in a frictionless market void of arbitrage opportunities and the number of potential discontinuities (jumps) in the price process per unit time is finite. Then the log-price process p is a semi-martingale (e. g., Back [33]) and therefore the cumulative return $r(t)$ admits the decomposition (e. g., Protter [229])

$$r(t) = \mu(t) + M^{\mathrm{C}}(t) + M^{\mathrm{J}}(t)\ , \tag{6}$$

where $\mu(t)$ is a predictable and finite variation process, $M^{\mathrm{C}}(t)$ a continuous-path infinite-variation martingale, and $M^{\mathrm{J}}(t)$ is a compensated finite activity jump martingale. Over a discrete time interval the decomposition (6) becomes

$$r(t, h) = \mu(t, h) + M^{\mathrm{C}}(t, h) + M^{\mathrm{J}}(t, h)\ , \tag{7}$$

where $\mu(t, h) = \mu(t) - \mu(t-h)$, $M^{\mathrm{C}}(t, h) = M^{\mathrm{C}}(t) - M^{\mathrm{C}}(t-h)$, and $M^{\mathrm{J}}(t, h) = M^{\mathrm{J}}(t) - M^{\mathrm{J}}(t-h)$.

Denote now with $[r, r]$ the quadratic variation of the semi-martingale process r, where (Protter [229])

$$[r, r]_t = r(t)^2 - 2 \int r(s-)\mathrm{d}r(s)\ , \tag{8}$$

and $r(t-) = \lim_{s \uparrow t} r(s)$. If the finite variation process μ is continuous, then its quadratic variation is identically zero and the predictable component μ in decomposition (7) does not affect the quadratic variation of the return r. Thus, we obtain an expression for the quadratic return variation over the time interval from $t-h$ to t, $0 \le h \le t \le T$ (e. g., Andersen et al. [21] and Barndorff-Nielsen and Shephard [51,52]):

$$\begin{aligned} \mathrm{QV}(t, h) &= [r, r]_t - [r, r]_{t-h} \\ &= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\ &\quad + \sum_{t-h<s\le t} \Delta M^2(s) \\ &= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\ &\quad + \sum_{t-h<s\le t} \Delta r^2(s)\ . \end{aligned} \tag{9}$$

Most continuous-time models for asset returns can be cast within the general setting of Eq. (7), and Eq. (9) provides a framework to study the model-implied return variance. For instance, the Black and Scholes [63] model is a special case of the setting described by Eq. (7) in which the conditional mean process μ is constant, the continuous martingale M^C is a standard Brownian motion process, and the jump martingale M^J is identically zero:

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sigma \mathrm{d}W(t)\ . \tag{10}$$

In this case, the quadratic return variation over the time interval from $t-h$ to t, $0 \le h \le t \le T$, simplifies to

$$\mathrm{QV}(t, h) = \int_{t-h}^{t} \sigma^2 \mathrm{d}s = \sigma^2 h\ , \tag{11}$$

that is, return volatility is constant over any time interval of length h.

A second notable example is the jump-diffusion model of Merton [214],

$$\mathrm{d}p(t) = (\mu - \lambda\bar{\xi})\mathrm{d}t + \sigma \mathrm{d}W(t) + \xi(t)\mathrm{d}q_t\ , \tag{12}$$

where q is a Poisson process uncorrelated with W and governed by the constant jump intensity λ, i. e., $\text{Prob}(\mathrm{d}q_t = 1) = \lambda \mathrm{d}t$. The scaling factor $\xi(t)$ denotes the magnitude of the jump in the return process if a jump occurs at time t. It is assumed to be normally distributed,

$$\xi(t) \rightsquigarrow N(\overline{\xi}, \sigma_{\xi}^2) . \tag{13}$$

In this case, the quadratic return variation process over the time interval from $t-h$ to t, $0 \leq h \leq t \leq T$ becomes

$$\begin{aligned} \text{QV}(t,h) &= \int_{t-h}^{t} \sigma^2 \mathrm{d}s + \sum_{t-h \leq s \leq t} J(s)^2 \\ &= \sigma^2 h + \sum_{t-h \leq s \leq t} J(s)^2 , \end{aligned} \tag{14}$$

where $J(t) \equiv \xi(t)\mathrm{d}q(t)$ is non-zero only if a jump actually occurs.

Finally, a broad class of stochastic volatility models is defined by

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t) + \xi(t)\mathrm{d}q_t , \tag{15}$$

where q is a constant-intensity Poisson process with log-normal jump amplitude (13). Equation (15) is also a special case of (7) and the associated quadratic return variation over the time interval from $t-h$ to t, $0 \leq h \leq t \leq T$, is

$$\begin{aligned} \text{QV}(t,h) &= \int_{t-h}^{t} \sigma(s)^2 \mathrm{d}s + \sum_{t-h \leq s \leq t} J(s)^2 \\ &\equiv IV(t,h) + \sum_{t-h \leq s \leq t} J(s)^2 . \end{aligned} \tag{16}$$

As in the general case of Eq. (9), Eq. (16) identifies the contribution of diffusive volatility, termed 'integrated variance' (IV), and cumulative squared jumps to the total quadratic variation.

Early applications typically ignored jumps and focused exclusively on the integrated variance component. For instance, IV plays a key role in Hull and White's [174] SV option pricing model, which we discuss in Sect. "Options" below along with other option pricing applications. For illustration, we focus here on the SV model specification by Wiggins [256]:

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sigma(t)\mathrm{d}W_p(t) \tag{17}$$

$$\mathrm{d}\sigma(t) = f(\sigma(t))\mathrm{d}t + \eta\sigma(t)\mathrm{d}W_{\sigma}(t) , \tag{18}$$

where the innovations to the return $\mathrm{d}p$ and volatility σ, W_p and W_{σ}, are standard Brownian motions. If we define $y = \log(\sigma)$ and apply Itô's formula we obtain

$$\begin{aligned} \mathrm{d}y(t) &= \mathrm{d}\log(\sigma(t)) \\ &= \left[-\frac{1}{2}\eta^2 + \frac{f(\sigma(t))}{\sigma(t)} \right] \mathrm{d}t + \eta \mathrm{d}W_{\sigma}(t) . \end{aligned} \tag{19}$$

Wiggins approximates the drift term $f(\sigma(t)) \approx \{\alpha + \kappa[\log(\overline{\sigma}) - \log(\sigma(t))]\}\sigma(t)$. Substitution in Eq. (19) yields

$$\mathrm{d}\log(\sigma(t)) = [\overline{\alpha} - \kappa \log(\sigma(t))]\mathrm{d}t + \eta \mathrm{d}W_{\sigma}(t) , \tag{20}$$

where $\overline{\alpha} = \alpha + \kappa \log(\overline{\sigma}) - \frac{1}{2}\eta^2$. As such, the logarithmic standard deviation process in Wiggins has diffusion dynamics similar in spirit to Taylor's discrete time AR(1) model for the logarithmic information process, Eq. (4). As in Taylor's model, negative correlation between return and volatility innovations, $\rho = \text{corr}(W_p, W_{\sigma}) < 0$, generates an asymmetric response of volatility to return shocks similar to the leverage effect in discrete-time EGARCH models.

More recently, several authors have imposed restrictions on the continuous-time SV jump-diffusion (15) that render the model more tractable while remaining consistent with the empirical features of the data. We return to these models in Sect. "Options" below.

Realized Volatility

Model-free measures of return variation constructed only from concurrent return realizations have been considered at least since Merton [215]. French et al. [148] construct monthly historical volatility estimates from daily return observations. More recently, the increased availability of transaction data has made it possible to refine early measures of historical volatility into the notion of 'realized volatility', which is endowed with a formal theoretical justification as an estimator of the quadratic return variation as first noted in Andersen and Bollerslev [18]. The realized volatility of an asset return r over the time interval from $t-h$ to t is

$$RV(t,h;n) = \sum_{i=1}^{n} r\left(t - h + \frac{ih}{n}, \frac{h}{n}\right)^2 . \tag{21}$$

Semi-martingale theory ensures that the realized volatility measure RV converges to the return quadratic variation QV, previously defined in Eq. (9), when the sampling frequency n increases. We point the interested reader to, e. g., Andersen et al. [19] to find formal arguments in support of this claim. Here we convey intuition for this result by considering the special case in which the asset return follows a continuous-time diffusion without jumps,

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t) . \tag{22}$$

As in Eq. (21), consider a partition of the $[t-h,t]$ interval with mesh h/n. A discretization of the diffusion (22) over a sub-interval from $(t-h+(i-1)h/n)$ to $(t-h+ih/n)$, $i=1,\dots,n$, yields

$$r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)\approx\mu\left(t-h+\frac{(i-1)h}{n}\right)\frac{h}{n}$$
$$+\sigma\left(t-h+\frac{(i-1)h}{n}\right)\Delta W\left(t-h+\frac{ih}{n}\right), \qquad (23)$$

where $\Delta W(t-h+ih/n)=W(t-h+ih/n)-W(t-h+(i-1)h/n)$.

Suppressing time indices, the squared return r^2 over the time interval of length h/n is therefore:

$$r^2=\mu^2\left(\frac{h}{n}\right)^2+2\mu\sigma\Delta W\left(\frac{h}{n}\right)+\sigma^2(\Delta W)^2\,. \qquad (24)$$

As $n\to\infty$ the first two terms vanish at a rate higher than the last one. In particular, to a first order approximation the squared return equals the squared return innovation and therefore the squared return conditional mean and variance are

$$\mathrm{E}\left[r^2|\mathcal{F}_t\right]\approx\sigma^2\frac{h}{n} \qquad (25)$$

$$\mathrm{Var}\left[r^2|\mathcal{F}_t\right]\approx 2\sigma^4\left(\frac{h}{n}\right)^2\,. \qquad (26)$$

The no-arbitrage condition implies that return innovations are serially uncorrelated. Thus, summing over $i=1,\dots,n$ we obtain

$$\mathrm{E}\left[RV(t,h,n)|\mathcal{F}_t\right]$$
$$=\sum_{i=1}^{n}\mathrm{E}\left[r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)^2|\mathcal{F}_t\right]$$
$$\approx\sum_{i=1}^{n}\sigma\left(t-h+\frac{(i-1)h}{n}\right)^2\frac{h}{n}$$
$$\approx\int_{t-h}^{t}\sigma(s)^2\mathrm{d}s \qquad (27)$$

$$\mathrm{Var}\left[RV(t,h,n)|\mathcal{F}_t\right]$$
$$=\sum_{i=1}^{n}\mathrm{Var}\left[r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)^2|\mathcal{F}_t\right]$$
$$\approx\sum_{i=1}^{n}2\sigma\left(t-h+\frac{(i-1)h}{n}\right)^4\left(\frac{h}{n}\right)^2$$
$$\approx 2\left(\frac{h}{n}\right)\int_{t-h}^{t}\sigma(s)^4\mathrm{d}s\,. \qquad (28)$$

Equation (27) illustrates that realized volatility is an unbiased estimator of the return quadratic variation, while Eq. (28) shows that the estimator is consistent as its variance shrinks to zero when we increase the sampling frequency n and keep the time interval h fixed. Taken together, these results suggest that RV is a powerful and model-free measure of the return quadratic variation. Effectively, RV gives practical empirical content to the latent volatility state variable underlying the models previously discussed in Sect. "Continuous-Time Stochastic Volatility Models".

Two issues complicate the practical application of the convergence results illustrated in Eqs. (27) and (28). First, a continuum of instantaneous return observations must be used for the conditional variance in Eq. (28) to vanish. In practice, only a discrete price record is observed, and thus an inevitable discretization error is present. Barndorff-Nielsen and Shephard [52] develop an asymptotic theory to assess the effect of this error on the RV estimate (see also [209]). Second, market microstructure effects (e. g., price discreteness, bid-ask spread positioning due to dealer inventory control, and bid-ask bounce) contaminate the return observations, especially at the ultra-high frequency. These effects tend to generate spurious correlations in the return series which can be partially eliminated by filtering the data prior to forming the RV estimates. However, this strategy is not a panacea and much current work studies the optimal sampling scheme and the construction of improved realized volatility in the presence of microstructure noise. This growing literature is surveyed by Hansen and Lunde [165], Bandi and Russell [46], McAleer and Medeiros [205], and Andersen and Benzoni [14]. Recent notable contributions to this literature include Bandi and Russell [45], Barndorff-Nielsen et al. [49], Diebold and Strasser [121], and Zhang, Mykland, and Aï t-Sahalia [262]. Related, there is the issue of how to construct RV measures when the market is rather illiquid. One approach is to use a lower sampling frequency and focus on longer-horizon RV measure. Alternatively the literature has explored volatility measures that are more robust to situations in which the noise-to-signal ratio is high, e. g., Alizadeh et al. [8], Brandt and Diebold [72], Brandt and Jones [73], Gallant et al. [151], Garman and Klass [157], Parkinson [221], Schwert [237], and Yang and Zhang [259] consider the high-low price range measure. Dobrev [122] generalizes the range estimator to high-frequency data and shows its link with RV measures.

Equations (27) and (28) also underscore an important difference between RV and other volatility measures. RV is an ex-post model-free estimate of the quadratic variation process. This is in contrast to ex-ante measures which attempt to forecast future quadratic variation using information up to current time. The latter class includes parametric GARCH-type volatility forecasts as well as forecasts built from stochastic volatility models through, e. g., the Kalman filter (e. g., [167,168]), the particle filter (e. g., [186,187]) or the reprojection method (e. g., [152,155]).

More recently, other studies have pursued more direct time-series modeling of volatility to obtain alternative ex-ante forecasts. For instance, Andersen et al. [21] follow an ARMA-style approach, extended to allow for long memory features, to model the logarithmic foreign exchange rate realized volatility. They find the fit to dominate that of traditional GARCH-type models estimated from daily data. In a related development, Andersen, Bollerslev, and Meddahi [24,25] exploit the general class of Eigenfunction Stochastic Volatility (ESV) models introduced by Meddahi [208] to provide optimal analytic forecast formulas for realized volatility as a function of past realized volatility. Other scholars have pursued more general model specifications to improve forecasting performance. Ghysels et al. [159] consider Mixed Data Sampling (MIDAS) regressions that use a combination of volatility measures estimates at different frequencies and horizons. Related, Engle and Gallo [137] exploit the information in different volatility measures, captured by a multivariate extension of the multiplicative error model suggested by Engle [136], to predict multi-step volatility. Finally, Andersen et al. [20] build on the Heterogeneous AutoRegressive (HAR) model by Barndorff-Nielsen and Shephard [50] and Corsi [110] and propose a HAR-RV component-based regression to forecast the h-steps ahead quadratic variation:

$$\begin{aligned}\mathrm{RV}(t+h,h) = \beta_0 + \beta_D \mathrm{RV}(t,1) + \beta_W \mathrm{RV}(t,5)\\ + \beta_M \mathrm{RV}(t,21) + \varepsilon(t+h)\,. \qquad (29)\end{aligned}$$

Here the lagged volatility components $\mathrm{RV}(t,1)$, $\mathrm{RV}(t,5)$, and $\mathrm{RV}(t,21)$ combine to provide a parsimonious approximation to the long-memory type behavior of the realized volatility series, which has been documented in several studies (e. g., Andersen et al. [19]). Simple OLS estimation yields consistent estimates for the coefficients in the regression (29), which can be used to forecast volatility out of sample.

As mentioned previously, the convergence results illustrated in Eqs. (27) and (28) stem from the theory of semi-martingales under conditions more general than those underlying the continuous-time diffusion in Eq. (22). For instance, these results are robust to the presence of discontinuities in the return path as in the jump-diffusion SV model (15). In this case the realized volatility measure (21) still converges to the return quadratic variation, which is now the sum of the diffusive integrated volatility IV and the cumulative squared jump component:

$$\mathrm{QV}(t,h) = \mathrm{IV}(t,h) + \sum_{t-h\le s\le t} J(s)^2\,. \qquad (30)$$

The decomposition in Eq. (30) motivates the quest for separate estimates of the two quadratic variation components, IV and squared jumps. This is a fruitful exercise in forecasting applications, since separate estimation of the two components increases predictive accuracy (e. g., [20]). Further, this decomposition is relevant for derivatives pricing, e. g., options are highly sensitive to jumps as well as large moves in volatility (e. g., [141,220]).

A consistent estimate of integrated volatility is the k-skip bipower variation, BV (e. g., Barndorff-Nielsen and Shephard [53]),

$$\begin{aligned}\mathrm{BV}(t,h;k,n) = \frac{\pi}{2}\sum_{i=k+1}^{n}\left|r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)\right|\\ \times\left|r\left(t-h+\frac{(i-k)h}{n},\frac{h}{n}\right)\right|\,. \qquad (31)\end{aligned}$$

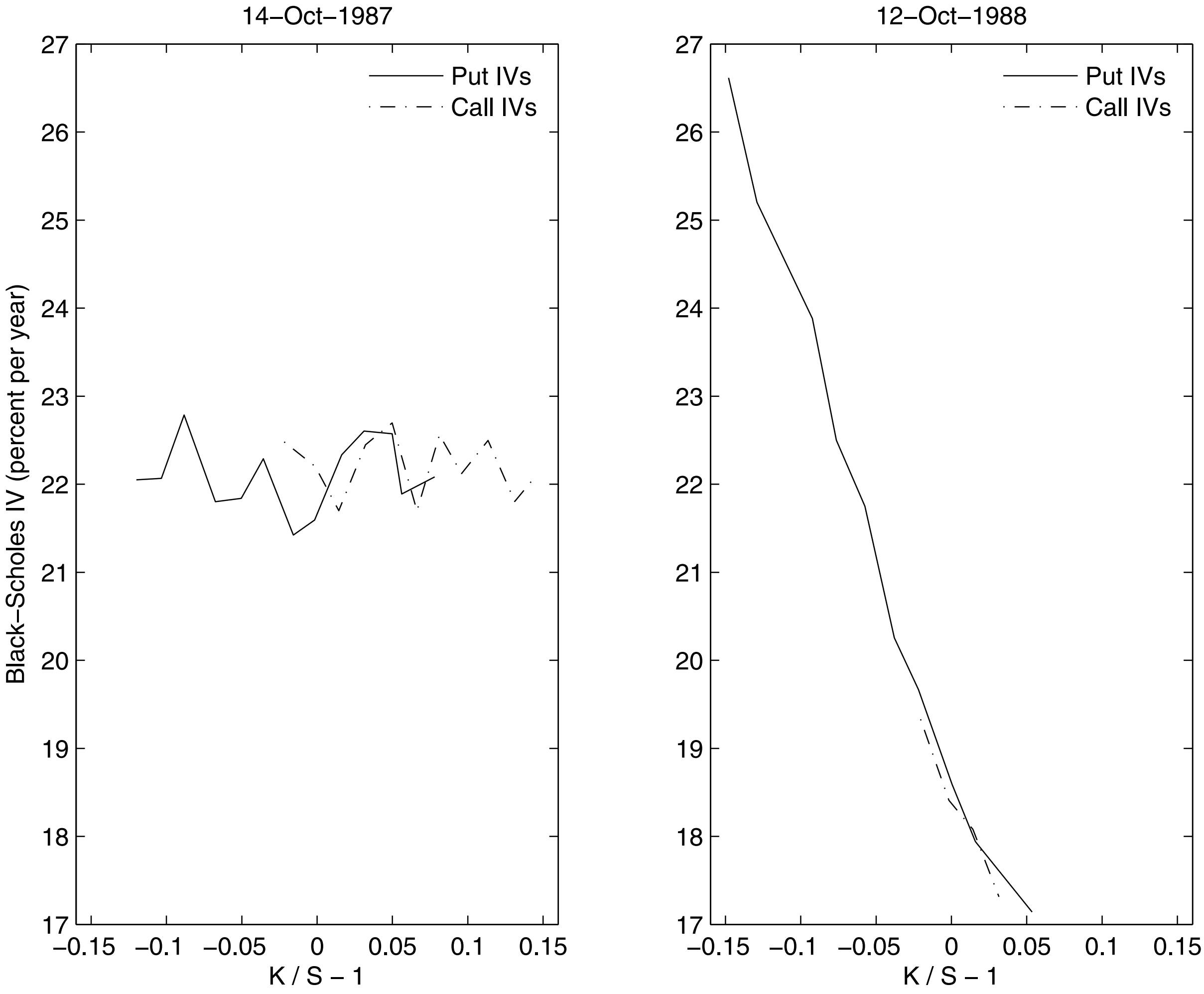
Liu and Maheu [202] and Forsberg and Ghysels [147] show that realized power variation, which is robust to the presence of jumps, can improve volatility forecasts. A well-known special case of (31) is the 'realized bipower variation', which has $k = 1$ and is denoted $\mathrm{BV}(t,h;n) \equiv \mathrm{BV}(t,h;1,n)$. We can combine bipower variation with the realized volatility RV to obtain a consistent estimate of the squared jump component, i. e.,

$$\begin{aligned}\mathrm{RV}(t,h;n) - \mathrm{BV}(t,h;n) \underset{n\to\infty}{\longrightarrow} \mathrm{QV}(t,h) - \mathrm{IV}(t,h)\\ = \sum_{t-h\le s\le t} J(s)^2\,. \qquad (32)\end{aligned}$$

The result in Eq. (32) are useful to design tests for the presence of jumps in volatility, e. g., Andersen et al. [20], Barndorff-Nielsen and Shephard [53,54], Huang and Tauchen [172], and Mizrach [217]. More recently, alternative approaches to test for jumps have been developed by Aït-Sahalia and Jacod [6], Andersen et al. [23], Lee and Mykland [195], and Zhang [261].

Applications

The power of the continuous-time paradigm has been evident ever since the work by Merton [212] on intertemporal portfolio choice, Black and Scholes [63] on option pricing, and Vasicek [255] on bond valuation. However, the idea of casting these problems in a continuous-time diffusion



Stochastic Volatility, Figure 1

Pre- and post-1987 crash implied volatilities. The plots depict Black-Scholes implied volatilities computed from near-maturity options on the S&P 500 futures on October 14, 1987 (the week before the 1987 market crash) and a year later

context goes back all the way to the work in 1900 by Bachelier [32].

Merton [213] develops a continuous-time general-equilibrium intertemporal asset pricing model which is later extended by Cox et al. [112] to a production economy. Because of its flexibility and analytical tractability, the Cox et al. [112] framework has become a key tool used in several financial applications, including the valuation of options and other derivative securities, the modeling of the term structure of risk-free interest rates, the pricing of foreign currencies and defaultable bonds.

Volatility has played a central role in these applications. For instance, an option's payoff is non-linear in the price of the underlying asset and this feature renders the option value highly sensitive to the volatility of underlying returns. Further, derivatives markets have grown rapidly in size and complexity and financial institutions have been facing the challenge to manage intricate portfolios exposed to multiple risk sources. Risk management of these sophisticated positions hinges on volatility modeling. More recently, the markets have responded to the increasing hedging demands of investors by offering a menu of new products including, e. g., volatility swaps and derivatives on implied volatility indices like the VIX. These innovations have spurred an even more pressing need to accurately measure and forecast volatility in financial markets.

Research has responded to these market developments. We next provide a brief illustrative overview of the recent literature dealing with option pricing and term structure modeling, with an emphasis on the role that volatility modeling has played in these two key applications.

Options

Rubinstein [233] and Bates [55], among others, note that prior to the 1987 market crash the Black and Scholes [63] (BS) formula priced option contracts quite accurately whereas after the crash it has been systematically underpricing out-of-the-money equity-index put contracts. This feature is evident from Fig. 1, which is constructed from options on the S&P 500 futures. It shows the implied volatility function for near-maturity contracts traded both before and after October 19, 1987 ('Black Monday'). The mild u-shaped pattern prevailing in the pre-crash implied volatilities is labeled a 'volatility smile,' in contrast to the asymmetric post-1987 'volatility smirk'. Importantly, while the steepness and level of the implied volatility curve fluctuate day to day depending on market conditions, the curve has been asymmetric and upward sloping ever since 1987, so the smirk remains in place to the current date, e. g., Benzoni et al. [60]. In contrast, before the crash the implied volatility curve was invariably flat or mildly u-shaped as documented in, e. g., [57]. Finally, we note that the post-1987 asymmetric smirk for index options contrasts sharply with the pattern for individual equity options, which possess flat or mildly u-shaped implied volatility curves (e. g., [37,65]).

Given the failures of the BS formula, much research has gone into relaxing the underlying assumptions. A natural starting point is to allow volatility to evolve randomly, inspiring numerous studies that examine the option pricing implications of SV models. The list of early contributions includes [174,188,211,238,244,245,256]. Here we focus in particular on the Hull and White [174] model,

$$\mathrm{d}p(t) = \mu_p \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) \tag{33}$$

$$\frac{\mathrm{d}V(t)}{V(t)} = \mu_V \mathrm{d}t + \sigma_V \mathrm{d}W_V(t)\,, \tag{34}$$

where W_p and W_V are standard Brownian motions. In general, shocks to returns and volatility may be (negatively) correlated, however for tractability Hull and White assume $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V) = 0$. Under this assumption they show that, in a risk-neutral world, the premium C^{HW} on a European call option is the Black and Scholes price C^{BS} evaluated at the average integrated variance $\overline{V}$,

$$\overline{V} = \frac{1}{T-t}\int_t^T V(s)\mathrm{d}s\,, \tag{35}$$

integrated over the distribution $h(\overline{V}|V(t))$ of $\overline{V}$:

$$C^{\mathrm{HW}}(p(t), V(t)) = \int C^{\mathrm{BS}}(\overline{V})h(\overline{V}|V(t))\mathrm{d}\overline{V}\,. \tag{36}$$

The early efforts to identify a more realistic probabilistic model for the underlying return were slowed by the analytical and computational complexity of the option pricing problem. Unlike the BS setting, the early SV specifications do not admit closed-form solutions. Thus, the evaluation of the option price requires time-consuming computations through, e. g., simulation methods or numerical solution of the pricing partial differential equation by finite difference methods. Further, the presence of a latent factor, volatility, and the lack of closed-form expressions for the likelihood function complicate the estimation problem.

Consequently, much effort has gone into developing restrictions for the distribution of the underlying return process that allow for (semi) closed-form solutions and are consistent with the empirical properties of the data. The 'affine' class of continuous-time models has proven particularly useful in providing a flexible, yet analytically tractable, setting. Roughly speaking, the defining feature of affine jump-diffusions is that the drift term, the conditional covariance term, and the jump intensity are all a linear-plus-constant (affine) function of the state vector. The Vasicek [255] bond valuation model and the Cox et al. [112] intertemporal asset pricing model provide powerful examples of the advantages of the affine paradigm.

To illustrate the progress in option pricing applications built on affine models, consider the return dynamics

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) + \xi_p(t)\mathrm{d}q(t) \tag{37}$$

$$\mathrm{d}V(t) = \kappa(\overline{V} - V(t))\mathrm{d}t + \sigma_V\sqrt{V(t)}\mathrm{d}W_V(t) + \xi_V(t)\mathrm{d}q(t)\,, \tag{38}$$

where W_p and W_V are standard Brownian motions with non-zero correlation $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V)$, q is a Poisson process, uncorrelated with W_p and W_V, with jump intensity

$$\lambda(t) = \lambda_0 + \lambda_1 V(t)\,, \tag{39}$$

that is, $\mathrm{Prob}(\mathrm{d}q_t = 1) = \lambda(t)\mathrm{d}t$. The jump amplitudes variables ξ_p and ξ_V have distributions

$$\xi_V(t) \rightsquigarrow \exp(\overline{\xi}_V) \tag{40}$$

$$\xi_p(t)|\xi_V(t) \rightsquigarrow N\left(\overline{\xi}_p + \rho_\xi \xi_V(t), \sigma_p^2\right)\,. \tag{41}$$

Here volatility is not only stochastic but also subject to jumps which occur simultaneously with jumps in the underlying return process. The Black and Scholes model is a special case of (37)–(41) for constant volatility, $V(t) = \sigma^2, 0 \leq t \leq T$, and no jumps, $\lambda(t) = 0, 0 \leq t \leq T$. The Merton [214] model arises from (37)–(41) if volatility is constant but we allow for jumps in returns.

More recently, Heston [170] has considered a special case of (37)–(41) with stochastic volatility but without jumps. Using transform methods he derives a European option pricing formula which may be evaluated readily through simple numerical integration. His SV model has GARCH-type features, in that the variance is persistent and mean reverts at a rate κ to the long-run mean $\overline{V}$. Compared to Hull and White's [174] setting, Heston's model allows for shocks to returns and volatility to be negatively correlated, i. e., $\rho < 0$, which creates a leverage-type effect and skews the return distribution. This feature is consistent with the properties of equity index returns. Further, a fatter left tail in the return distribution results in a higher cost for crash insurance and therefore makes out-of-the-money put options more expensive. This is qualitatively consistent with the patterns in implied volatilities observed after the 1987 market crash and discussed above.

Bates [56] has subsequently extended Heston's approach to allow for jumps in returns and using similar transform methods he has obtained a semi-closed form solution for the option price. The addition of jumps provides a more realistic description of equity returns and has important option pricing implications. With diffusive shocks (e. g., stochastic volatility) alone a large drop in the value of the underlying asset over a short time span is very unlikely whereas a market crash is always possible as long as large negative jumps can occur. This feature increases the value of a short-dated put option, which offers downside protection to a long position in the underlying asset.

Finally, Duffie et al. [130] have introduced a general model with jumps to volatility which embeds the dynamics (37)–(41). In model (37)–(41), the likelihood of a jump to occur increases when volatility is high ($\lambda_1 > 0$) and a jump in returns is accompanied by an outburst of volatility. This is consistent with what is typically observed during times of market stress. As in the Heston case, variance is persistent with a mean reversion coefficient κ towards its *diffusive* long-run mean $\overline{V}$, while the total long-run variance mean is the sum of the diffusive and jump components. In the special case of constant jump intensity, i. e., $\lambda_1 = 0$, the total long-run mean is $\overline{V} + \overline{\xi}_V \lambda_0 / \kappa$. The jump term $(\xi_V(t) \mathrm{d}q(t))$ fattens the right tail of the variance distribution, which induces leptokurtosis in the return distribution. Two effects generate asymmetrically distributed returns. The first channel is the diffusive leverage effect, i. e., $\rho < 0$, the second is the correlation between the volatility and the jump amplitude of returns generated through the coefficient ρ_ξ. Taken together, these effects increase model-implied option prices and help produce a realistic volatility smirk.

Several empirical studies rely on models of the form (37)–(41) in option-pricing applications. For instance, Bates [56] uses Deutsche Mark options to estimate a model with stochastic volatility and constant-intensity jumps to returns, while Bates [57] fits a jump-diffusion model with two SV factors to options on S&P 500 futures. In the latter case, the two SV factors combine to help capture features of the long-run memory in volatility while retaining the analytical tractability of the affine setting (see, e. g., [101] for another model with similar features). Alternative approaches to model long memory in continuous-time SV models rely on the fractional Brownian motion process, e. g., Comte and Renault [108] and Comte et al. [107], while Breidt et al. [76], Harvey [166] and Deo et al. [118] consider discrete-time SV models (see [175] for a review). Bakshi et al. [34,37] estimate a model similar to the one introduced by Bates [56] using S&P 500 options.

Other scholars rely on underlying asset return data alone for estimation. For instance, Andersen et al. [15] and Chernov et al. [95] use equity-index returns to estimate jump-diffusion SV models within and outside the affine (37)–(41) class. Eraker et al. [142] extend this analysis and fit a model that includes constant-intensity jumps to returns and volatility.

Finally, another stream of work examines the empirical implications of SV jump-diffusions using a joint sample of S&P 500 options and index returns. For example, Benzoni [59], Chernov and Ghysels [93], and Jones [190] estimate different flavors of the SV model without jumps. Pan [220] fits a model that has jumps in returns with time-varying intensity, while Eraker [141] extends Pan's work by adding jumps in volatility.

Overall, this literature has established that the SV jump-diffusion model dramatically improves the fit of underlying index returns and options prices compared to the Black and Scholes model. Stochastic volatility alone has a first-order effect and jumps further enhance model performance by generating fatter tails in the return distribution and reducing the pricing error for short-dated options. The benefits of the SV setting are also significant in hedging applications.

Another aspect related to the specification of SV models concerns the pricing of volatility and jump risks.

Stochastic volatility and jumps are sources of uncertainty. It is an empirical issue to determine whether investors demand to be compensated for bearing such risks and, if so, what the magnitude of the risk premium is. To examine this issue it is useful to write model (37)–(41) in so-called risk-neutral form. It is common to assume that the volatility risk premium is proportional to the instantaneous variance, $\eta(t) = \eta_V V(t)$. Further, the adjustment for jump risk is accomplished by assuming that the amplitude $\tilde{\xi}_p(t)$ of jumps to returns has mean $\bar{\tilde{\xi}}_p = \bar{\xi}_p + \eta_p$. These specifications are consistent with an arbitrage-free economy. More general specifications can also be supported in a general equilibrium setting, e. g., a risk adjustment may apply to the jump intensity $\lambda(t)$. However, the coefficients associated to these risk adjustments are difficult to estimate and to facilitate identification they typically are fixed at zero. Incorporating such risk premia in model (37)–(41) yields the following risk-neutral return dynamics (e. g., Pan [220] and Eraker [141]):

$$\mathrm{d}p(t) = (r - \mu^*)\mathrm{d}t + \sqrt{V(t)}\mathrm{d}\widetilde{W}_p(t) + \tilde{\xi}_p(t)\mathrm{d}q(t) \quad (42)$$

$$\begin{aligned} \mathrm{d}V(t) = {} & [\kappa(\overline{V} - V(t)) + \eta_V V(t)]\mathrm{d}t \\ & + \sigma_V \sqrt{V(t)}\mathrm{d}\widetilde{W}_V(t) + \xi_V(t)\mathrm{d}q(t) \,, \end{aligned} \quad (43)$$

where r is the risk-free rate, μ^* a jump compensator term, $\widetilde{W}_p$ and $\widetilde{W}_V$ are standard Brownian motions under this so-called $\mathcal{Q}$ measure, and the risk-adjusted jump amplitude variable $\tilde{\xi}_p$ is assumed to follow the distribution,

$$\tilde{\xi}_p(t)|\xi_V(t) \rightsquigarrow N\left(\bar{\tilde{\xi}}_p + \rho_\xi \xi_V(t), \sigma_p^2\right) \,. \quad (44)$$

Several studies estimate the risk-adjustment coefficients η_V and η_p for different specifications of model (37)–(44); see, e. g., Benzoni [59], Broadie et al. [78], Chernov and Ghysels [93], Eraker [141], Jones [190], and Pan [220]. It is found that investors demand compensation for volatility and jump risks and these risk premia are important for the pricing of index options. This evidence is reinforced by other studies examining the pricing of volatility risk using less structured but equally compelling procedures. For instance, Coval and Shumway [111] find that the returns on zero-beta index option straddles (i. e., combinations of calls and puts that have offsetting covariances with the index) are significantly lower than the risk-free return. This evidence suggests that in addition to market risk at least a second factor (likely, volatility) is priced in the index option market. Similar conclusions are obtained by Bakshi and Kapadia [36], Buraschi and Jackwerth [79], and Broadie et al. [78].

Risk-Free Bonds and Their Derivatives

The market for (essentially) risk-free Treasury bonds is liquid across a wide maturity spectrum. No-arbitrage restrictions constrain the allowable dynamics in the cross-section of bond yields. Much work has gone into the development of tractable dynamic term structure models capable of capturing the salient time-series properties of interest rates while respecting such cross-sectional no-arbitrage conditions. The class of so-called 'affine' dynamic term structure models provides a flexible and arbitrage-free, yet analytically tractable, setting for capturing the dynamics of the term structure of interest rates. Following Duffie and Kan [129], Dai and Singleton [114,115], and Piazzesi [226], the short term interest rate, $y_0(t)$, is an affine (i. e., linear-plus-constant) function of a vector of state variables, $X(t) = \{x_i(t),\ i = 1, \ldots, N\}$:

$$y_0(t) = \delta_0 + \sum_{i=1}^{N} \delta_i x_i(t) = \delta_0 + \delta_X' X(t) \,, \quad (45)$$

where the state-vector X has risk-neutral dynamics

$$\mathrm{d}X(t) = \tilde{\mathcal{K}}(\tilde{\Theta} - X(t))\mathrm{d}t + \Sigma\sqrt{S(t)}\mathrm{d}\widetilde{W}(t) \,. \quad (46)$$

In Eq. (46), $\widetilde{W}$ is an N-dimensional Brownian motion under the so-called $\mathcal{Q}$-measure, $\tilde{\mathcal{K}}$ and $\tilde{\Theta}$ are $N \times N$ matrices, and $S(t)$ is a diagonal matrix with the ith diagonal element given by $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Within this setting, the time-t price of a zero-coupon bond with time-to-maturity τ is given by

$$P(t, \tau) = \mathrm{e}^{A(\tau) - B(\tau)' X(t)} \,, \quad (47)$$

where the functions $A(\tau)$ and $B(\tau)$ solve a system of ordinary differential equations (ODEs); see, e. g., Duffie and Kan [129]. Semi-closed form solutions are also available for bond derivatives, e. g., bond options as well as caps and floors (e. g., Duffie et al. [130]).

In empirical applications it is important to also establish the evolution of the state vector X under the physical probability measure $\mathcal{P}$, which is linked to the $\mathcal{Q}$-dynamics (46) through a market price of risk, $\Lambda(t)$. Following Dai and Singleton [114] the market price of risk is often given by

$$\Lambda(t) = \sqrt{S(t)}\lambda \,, \quad (48)$$

where λ is an $N \times 1$ vector of constants. More recently, Duffee [127] proposed a broader 'essentially affine' class, which retains the tractability of standard models but, in contrast to the specification in Eq. (48), allows compensation for interest rate risk to vary independently of interest rate volatility. This additional flexibility proves useful in forecasting future yields. Subsequent generalization are in Duarte [124] and Cheridito et al. [92].

Litterman and Scheinkman [201] demonstrate that virtually all variation in US Treasury rates is captured by three factors, interpreted as changes in 'level', 'steepness', and 'curvature'. Consistent with this evidence, much of the term-structure literature has focused on three-factor models. One problem with these models, however, is that the factors are latent variables void of immediate economic interpretation. As such, it is challenging to impose appropriate identifying conditions for the model coefficients and in particular to find the ideal representation for the 'most flexible' model, i. e., the model with the highest number of identifiable coefficients. Dai and Singleton [114] conduct an extensive specification analysis of multi-factor affine term structure models. They classify these models into subfamilies according to the number of (independent linear combination of) state variables that determine the conditional variance matrix of the state vector. Within each subfamily, they proceed to identify the models that lead to well-defined bond prices (a condition they label 'admissibility') and among the admissible specifications they identify a 'maximal' model that nests econometrically all others in the subfamily. Joslin [191] builds on Dai and Singleton's [114] work by pursuing identification through a normalization of the drift term in the state vector dynamics (instead of the diffusion term, as in Dai and Singleton [114]). Duffie and Kan [129] follow an alternative approach to obtain an identifiable model by rotating from a set of latent state variables to a set of observable zero-coupon yields. Collin-Dufresne et al. [104] build on the insights of both Dai and Singleton [114] and Duffie and Kan [129]. They perform a rotation of the state vector into a vector that contains the first few components in the Taylor series expansion of the yield curve around a maturity of zero and their quadratic variation. One advantage is that the elements of the rotated state vector have an intuitive and unique economic interpretation (such as level, slope, and curvature of the yield curve) and therefore the model coefficients in this representation are identifiable. Further, it is easy to construct a model-independent proxy for the rotated state vector, which facilitates model estimation as well as interpretation of the estimated coefficients across models and sample periods.

This discussion underscores an important feature of affine term structure models. The dependence of the conditional factor variance $S(t)$ on one or more of the elements in X introduces stochastic volatility in the yields. However, when a square-root factor is present parametric restrictions (admissibility conditions) need to be imposed so that the conditional variance $S(t)$ is positive over the range of X. These restrictions affect the correlations among the factors which, in turn, tend to worsen the cross-sectional fit of the model. Specifically, CIR models in which $S(t)$ depends on all the elements of X require the conditional correlation among the factors to be zero, while the admissibility conditions imposed on the matrix $\mathcal{K}$ renders the unconditional correlations non-negative. These restrictions are not supported by the data. In contrast, constant-volatility Gaussian models with no square-root factors do not restrict the signs and magnitude of the conditional and unconditional correlations among the factors but they do, of course, not accommodate the pronounced and persistent volatility fluctuations observed in bond yields. The class of models introduced by Dai and Singleton [114] falls between these two extremes. By including both Gaussian and square-root factors they allow for time-varying conditional volatilities of the state variables and yet they do not constrain the signs of some of their correlations. This flexibility helps to address the trade off between generating realistic correlations among the factors while capturing the time-series properties of the yields' volatility.

A related aspect of (unconstrained) affine models concerns the dual role that square-root factors play in driving the time-series properties of yields' volatility and the term structure of yields. Specifically, the time-t yield $y_\tau(t)$ on a zero-coupon bond with time-to-maturity τ is given by

$$P(t,\tau) = \mathrm{e}^{-\tau y_\tau(t)} \,. \tag{49}$$

Thus, we have

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) \,. \tag{50}$$

It is typically assumed that the B matrix has full rank and therefore Eq. (50) provides a direct link between the state-vector $X(t)$ and the term-structure of bond yields. Further, Itô's Lemma implies that the yield y_τ also follows a diffusion process:

$$\mathrm{d}y_\tau(t) = \mu_{y_\tau}(X(t),t)\mathrm{d}t + \frac{B(\tau)'}{\tau} \Sigma \sqrt{S(t)}\mathrm{d}\widetilde{W}(t) \,. \tag{51}$$

Consequently, the (instantaneous) quadratic variation of the yield given as the squared yield volatility coefficient for y_τ is

$$V_{y_\tau}(t) = \frac{B(\tau)'}{\tau} \Sigma S(t) \Sigma' \frac{B(\tau)}{\tau} \,. \tag{52}$$

The elements of the $S(t)$ matrix are affine in the state vector $X(t)$, i. e., $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Further, invoking the full rank condition on $B(\tau)$, Eq. (50) implies that each state variable in the vector $X(t)$ is an affine function of the bond yields $Y(t) = \{y_{\tau_j}(t),\ j = 1, \ldots, J\}$. Thus, for any τ there is a set of constants $a_{\tau,j},\ j = 0, \ldots, J$, so that

$$V_{y_\tau}(t) = a_{\tau,0} + \sum_{j=1}^{J} a_{\tau,j} y_{\tau_j}(t)\,. \tag{53}$$

Hence, the current quadratic yield variation for bonds at any maturity is a linear combination of the term structure of yields. As such, the market is complete, i. e., volatility is perfectly spanned by a portfolio of bonds.

Collin-Dufresne and Goldstein [103] note that this spanning condition is unnecessarily restrictive and propose conditions which ensures that volatility no longer directly enters the main bond pricing Eq. (47). This restriction, which they term ‘unspanned stochastic volatility’ (USV), effectively breaks the link between the yields’ quadratic variation and the level of the term structure by imposing a reduced rank condition on the $B(\tau)$ matrix. Further, since their model is a special (nested) case of the affine class it retains the analytical tractability of the affine model class. Recently Joslin [191] has derived more general conditions for affine term structure models to exhibit USV. His restrictions also produce a market incompleteness (i. e., volatility cannot be hedged using a portfolio of bonds) but do not constrain the degree of mean reversion of the other state variables so that his specification allows for more flexibility in capturing the persistence in interest rate series. (See also the USV conditions in the work by Trolle and Schwartz [253]).

There is conflicting evidence on the volatility spanning condition in fixed income markets. Collin-Dufresne and Goldstein [103] find that swap rates have limited explanatory power for returns on at-the-money ‘straddles’, i. e., portfolios mainly exposed to volatility risk. Similar findings are in Heidari and Wu [169], who show that the common factors in LIBOR and swap rates explain only a limited part of the variation in the swaption implied volatilities. Moreover, Li and Zhao [197] conclude that some of the most sophisticated multi-factor dynamic term structure models have serious difficulties in hedging caps and cap straddles, even though they capture bond yields well. In contrast, Fan et al. [143] argue that swaptions and even swaption straddles can be well hedged with LIBOR bonds alone, supporting the notion that bond markets are complete.

More recently other studies have examined several versions of the USV restriction, again coming to different conclusions. A direct comparison of these results, however, is complicated by differences in the model specification, the estimation method, and the data and sample period used in the estimation. Collin-Dufresne et al. [105] consider swap rates data and fit the model using a Bayesian Markov Chain Monte Carlo method. They find that a standard three-factor model generates a time series for the variance state variable that is essentially unrelated to GARCH estimates of the quadratic variation of the spot rate process or to implied variances from options, while a four-factor USV model generates both realistic volatility estimates and a good cross-sectional fit. In contrast, Jacobs and Karoui [178] consider a longer data set of US Treasury yields and pursue quasi-maximum likelihood estimation. They find the correlation between model-implied and GARCH volatility estimates to be high. However, when estimating the model with a shorter sample of swap rates, they find such correlations to be small or negative. Thompson [250] explicitly tests the Collin-Dufresne and Goldstein [103] USV restriction and rejects it using swap rates data. Bikbov and Chernov [62], Han [164], Jarrow et al. [183], Joslin [192], and Trolle and Schwartz [254] rely on data sets of derivatives prices and underlying interest rates to better identify the volatility dynamics.

Andersen and Benzoni [12] directly relate model-free realized volatility measures (constructed from high-frequency US Treasury data) to the cross-section of contemporaneous bond yields. They find that the explanatory power of such regressions is very limited, which indicates that volatility is not spanned by a portfolio of bonds. The evidence in Andersen and Benzoni [12] is consistent with the USV models of Collin-Dufresne et al. [105] and Joslin [191], as well as with a model that embeds weak dependence between the yields and volatility as in Joslin [192]. Moreover, Duarte [125] argues that the effects of mortgage-backed security hedging activity affects both the interest rate volatility implied by options and the actual interest rate volatility. This evidence suggests that variables that are not in the span of the term structure of yields and forward rates contribute to explain volatility in fixed income markets. Also related, Wright and Zhou [258] find that adding a measure of market jump volatility risk to a regression of excess bond returns on the term structure of forward rates nearly doubles the R^2 of the regression. Taken together, these findings suggest more generally that genuine SV models are critical for appropriately capturing the dynamic evolution of the term structure.

Estimation Methods

There are a very large number of alternative approaches to estimation and inference for parametric SV models and we abstain from a thorough review. Instead, we point to the basic challenges that exist for different types of specifications, how some of these were addressed in the early literature and finally provide examples of methods that have been used extensively in recent years. Our exposition continues to focus on applications to equity returns, interest rates, and associated derivatives.

Many of the original SV models were cast in discrete time, inspired by the popular GARCH paradigm. In that case, the distinct challenge for SV models is the presence of a strongly persistent latent state variable. However, more theoretically oriented models, focusing on derivatives applications, were often formulated in continuous time. Hence, it is natural that the econometrically-oriented literature has moved in this direction in recent years as well. This development provides an added complication as the continuous-time parameters must be estimated from discrete return data and without direct observations on volatility. For illustration, consider a fully parametric continuous-time SV model for the asset return r with conditional variance V and coefficient vector Ψ. Most methods to estimate Ψ rely on the conditional density f for the data generating process,

$$f(r(t), V(t)|\mathcal{I}(t-1), \Psi) = f_{r|V}(r(t)|V(t), \mathcal{I}(t-1), \Psi) \times f_V(V(t)|\mathcal{I}(t-1), \Psi) , \quad (54)$$

where $\mathcal{I}(t-1)$ is the available information set at time $t-1$. The main complications are readily identified. First, analytic expressions for the discrete-time transition (conditional) density, f, or the discrete-time moments implied by the data generating process operating in continuous time, are often unavailable. Second, volatility is latent in SV models, so that even if a closed-form expression for f is known, direct evaluation of the above expression is infeasible due to the absence of explicit volatility measures. The marginal likelihood with respect to the observable return process alone is obtained by integrating over all possible paths for the volatility process, but this integral has a dimension corresponding to sample size, rendering the approach infeasible in general.

Similar issues are present when estimating continuous-time dynamic term structure models. Following Piazzesi [227], a change of variable gives the conditional density for a zero-coupon yield y on a bond with time to maturity τ:

$$f(y_\tau(t)|\mathcal{I}(t-1), \Psi) = f_X(g(y_\tau(t), \Psi)|\mathcal{I}(t-1), \Psi) \times |\nabla_y g(y_\tau(t), \Psi)| . \quad (55)$$

Here the latent state vector X has conditional density f_X, the function $g(\cdot, \Psi)$ maps the observable yield y into X, $X(t) = g(y_\tau(t), \Psi)$, and $\nabla_y g(y_\tau(t), \Psi)$ is the Jacobian determinant of the transformation. Unfortunately, analytic expressions for the conditional density f_X are known only in some special cases. Further, the mapping $X(t) = g(y_\tau(t), \Psi)$ holds only if the model provides an exact fit to the yields, while in practice different sources of error (e. g., model mis-specification, microstructure effects, measurement errors) inject a considerable degree of noise into this otherwise deterministic linkage (for correct model specification) between the state vector and the yields. As such, a good measure of X might not be available to evaluate the conditional density (55).

Estimation via Discrete-Time Model Specification or Approximation

The first empirical studies have estimated discrete-time SV models via a (Generalized) Method of Moments procedure by matching a number of theoretical and sample moments, e. g., Chan et al. [89], Ho et al. [171], Longstaff and Schwartz [204], and Melino and Turnbull [211]. These models were either explicitly cast in discrete time or were seen as approximate versions of the continuous-time process of interest. Similarly, several authors estimate diffusive affine dynamic term structure models by approximating the continuous-time dynamics with a discrete-time process. If the error terms are stipulated to be normally distributed, the transition density of the discretized process is multivariate normal and computation of unconditional moments then only requires knowledge of the first two moments of the state vector. This result facilitates quasi-maximum likelihood estimation. In evaluating the likelihood function, some studies suggest using closed-form expressions for the first two moments of the continuous-time process instead of the moments of the discretized process (e. g., Fisher and Gilles [145] and Duffee [127]), thus avoiding the associated discretization bias. This approach typically requires some knowledge of the state of the system which may be obtained, imperfectly, through matching the system, given the estimated parameter vector, to a set of observed zero-coupon yields to infer the state vector X. A modern alternative is to use the so-called particle filter as an efficient filtering procedure for the unobserved state variables given the estimated parameter vector. We provide more detailed accounts of both of these procedures later in this section.

Finally, a number of authors develop a simulated maximum likelihood method that exploit the specific structure of the discrete-time SV model. Early examples are Danielsson and Richard [117] and Danielsson [116] who exploit the Accelerated Gaussian Importance Sampler for efficient

Monte Carlo evaluation of the likelihood. Subsequent improvements were provided by Fridman and Harris [149] and Liesenfeld and Richard [200], with the latter relying on Efficient Importance Sampling (EIS). In a second step, EIS can also be used for filtering the latent volatility state vector. In general, these inference techniques provide quite impressive efficiency but the methodology is not always easy to generalize beyond the structure of the basic discrete-time SV asset return model. We discuss the general inference problem for continuous-time SV models for which the lack of a closed-form expression for the transition density is an additional complicating factor in a later section.

Filtering the Latent State Variable Directly During Estimation

Some early studies focused on direct ways to extract estimates of the latent volatility state variable in discrete-time SV asset return models. The initial approach was based on quasi-maximum likelihood (QML) methods exploiting the Kalman filter. This method requires a transformation of the SV model to a linear state-space form. For instance, Harvey and Shephard [168] consider a version of the Taylor's [249] discrete-time SV model,

$$p(t) = p(t-1) + \beta + \sqrt{V(t)}\varepsilon(t) \tag{56}$$

$$\log(V(t)) = \alpha + \phi \log(V(t-1)) + \eta(t) , \tag{57}$$

where p is the logarithmic price, ε is a zero-mean error term with unit variance, and η is an independently-distributed error term with zero mean and variance σ_η^2.

Define $y(t) = p(t) - p(t-1) - \beta$, square the observations in Eq. (56), and take logarithms to obtain the *measurement equation*,

$$\ell(t) = \omega + h(t) + \xi(t) , \tag{58}$$

where $\ell(t) \equiv \log y(t)^2$, $h(t) \equiv \log(V(t))$. Further, ξ is a zero-mean disturbance term given by $\xi(t) = \log(\varepsilon(t)^2) - \mathrm{E}[\log(\varepsilon(t)^2)]$, $\omega = \log(\sigma^2) + \mathrm{E}[\log(\varepsilon(t)^2)]$, and σ is a scale constant which subsumes the effect of the drift term α in Eq. (57). The autoregression (57) yields the *transition equation*,

$$h(t) = \phi h(t-1) + \eta(t) , \tag{59}$$

Taken together, Eqs. (58) and (59) are the linear state-space transformation of the SV model (56)–(57). If the joint distribution of ε and η is symmetric, i. e., $f(\varepsilon, \eta) = f(-\varepsilon, -\eta)$, then the disturbance terms in the state-space form are uncorrelated even if η and ε are not. A possible dependence between ε and η allows the model to pick up some of the asymmetric behavior often observed in stock returns. Projection of $[h(t) - \mathrm{E}_{t-1}\, h(t)]$ over $[\,\ell(t) - \mathrm{E}_{t-1}\,\ell(t)\,]$ yields the Kalman filter estimate of the latent (logarithmic) variance process:

$$\begin{aligned}\mathrm{E}_t\, h(t) = \mathrm{E}_{t-1}\, h(t) &\\ &+ \frac{\mathrm{E}\{[h(t) - \mathrm{E}_{t-1}\, h(t)] \times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]\}}{\mathrm{E}\{[\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]^2\}}\\ &\times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)] ,\end{aligned} \tag{60}$$

where the conditional expectations $\mathrm{E}_{t-1}\,\ell(t)$ and $\mathrm{E}_{t-1}\, h(t)$ are given by:

$$\mathrm{E}_{t-1}\,\ell(t) = \omega + \mathrm{E}_{t-1}\, h(t) \tag{61}$$

$$\mathrm{E}_{t-1}\, h(t) = \phi\, \mathrm{E}_{t-1}\, h(t-1) . \tag{62}$$

To start the recursion (60)–(62), the initial value $\mathrm{E}_0\, h(0)$ is fixed at the long-run mean $\log(\overline{V})$.

Harvey and Shephard [168] estimate the model coefficients via quasi-maximum likelihood, i. e. by treating the errors ξ and η as though they were normal and maximizing the prediction-error decomposition form of the likelihood function obtained via the Kalman filter. Inference is valid as long as the standard errors are appropriately adjusted. In their application they rely on daily returns on the value-weighted US market index over 1967–1987 and daily returns for 30 individual stocks over 1974–1983. Harvey et al. [167] pursue a similar approach to fit a multivariate SV model to a sample of four exchange rate series from 1981 to 1985. One major drawback of the Kalman filter approach is that the finite sample properties can be quite poor because the error term, ξ, is highly non-Gaussian, see, e. g., Andersen, Chung, and Sørensen [27]. The method may be extended to accommodate various generalizations including long memory persistence in volatility as detailed in Ghysels, Harvey, and Renault [158].

A related literature, often exploited in multivariate settings, specifies latent GARCH-style dynamics for a state vector which governs the systematic evolution of a higher dimensional set of asset returns. An early representative of these specifications is in Diebold and Nerlove [120], who exploit the Kalman filter for estimation, while Fiorentina et al. [144] provide a likelihood-based estimation procedure using MCMC techniques. We later review the MCMC approach and the associated filtering application, e. g, the 'particle filter', in some detail.

The state-space form is also useful to characterize the dynamics of interest rates. Following, e. g., Piazzesi [226], for a discrete-time dynamic term structure model the measurement and transition equations are

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) + \xi_\tau(t) \tag{63}$$

$$X(t) = \mu + \Phi X(t-1) + \Sigma \sqrt{S(t)}\, \varepsilon(t)\,, \tag{64}$$

where $S(t)$ is a matrix whose elements are affine functions of the state vector X, and A and B solve a system of difference equations. When all the yields are observed with error (i. e., $\xi_\tau \neq 0 \forall \tau,\ 0 \leq \tau \leq T$), QML estimation of the system (63)–(64) via the extended Kalman filter method yields an estimate of the coefficient vector. Applications of this approach for the US term structure data include Campbell and Viceira [81], Gong and Remolona [161], and Pennacchi [225]. The extended Kalman filter involves a linear approximation of the relation between the observed data and the state variables, and the associated approximation error will produce biased estimates. Christoffersen et al. [99] raise this concern and recommend the use of the so-called unscented Kalman filter for estimation of systems in which the relation between data and state variables is highly non-linear, e. g., options data.

Methods Accommodating the Lack of a Closed-Form Transition Density

We have so far mostly discussed estimation techniques for models with either a known transition density or one that is approximated by a discrete-time system. However, the majority of empirically-relevant continuous-time models do not possess explicit transition densities and alternative approaches are necessary. This problem leads us naturally towards the large statistics and econometric literature on estimation of diffusions from discretely-observed data. The vast majority of these studies assume that all relevant variables are observed so the latent volatility or yield curve state variables, integral to SV models, are not accounted for. Nonetheless, it may be feasible to extract the requisite estimates of the state variable by alternate means, thus restoring the feasibility, albeit not efficiency, of the basic approach. Since the literature is large and not directly geared towards genuine SV models, we focus on methods that have seen use in applications involving latent state variables.

A popular approach is to invert the map between the state vector and a subset of the observables assuming that the model prices specific securities exactly. In applications to equity markets this is done, e. g., by assuming that one option contract is priced without error, which implies a specific value (estimate) of the variance process given the model parameters Ψ. For instance, Pan [220] follows this approach in her study of S&P 500 options and returns, which we review in more detail in Sect. "Estimation from Option Data". In applications to fixed income markets it is likewise stipulated that certain bonds are priced without error, i. e., in Eq. (63) the error term $\xi_{\tau_i}(t)$ is fixed at zero for a set of maturities $\tau_1, \ldots, \tau_N$, where N matches the dimension of the state vector X. This approach yields an estimate for the latent variables through the inverse-map $X(t) = g(y_\tau(t), \Psi)$.

One criticism of the state vector inversion procedure is that it requires ad hoc assumptions regarding the choice of the securities that are error-free (those used to compute model-implied measures of the state vector) vis-a-vis those observed with error (used either for estimation or to assess model performance in an 'out-of-sample' cross-sectional check). In fact, the extracted state vector can be quite sensitive to the choice of derivatives (or yields) used. Nevertheless, this approach has intuitive appeal. Model-implied measures of the state vector, in combination with a closed-form expression for the conditional density (55), allow for efficient estimation of the coefficient vector Ψ via maximum likelihood. Analytic expressions for f_X in Eq. (55) exist in a limited number of cases. For instance, if X is Gaussian then f_X is multivariate normal, while if X follows a square-root process then f_X can be expressed in terms of the modified Bessel function (e. g., [113]). Different flavors of these continuous-time models are estimated in, e. g., [91,106,132,182,223]. In more general cases, including affine processes that combine Gaussian and square-root state variables, closed-form expressions for f_X are no longer available. In the rest of this section we briefly review different methods to overcome this problem. The interested reader may consult, e. g., [226] for more details.

Lo [203] warns that the common approach of estimating parameters of an Itô process by applying maximum likelihood to a discretization of the stochastic differential equation yields inconsistent estimators. In contrast, he characterizes the likelihood function as a solution to a partial differential equation. The method is very general, e. g., it applies not only to continuous-time diffusions but also to jump processes. In practice, however, analytic solutions to the partial differential equations (via, e. g., Fourier transforms) are available only for a small class of models so computationally-intensive methods (e. g., finite differencing or simulations) are generally required to solve the problem. This is a severe limitation in the case of multivariate systems like SV models.

For general Markov processes, where the above solution is infeasible, a variety of procedures have been advocated in recent years. Three excellent surveys provide different perspectives on the issue. Aït-Sahalia, Hansen, and Scheinkman [5] discuss operator methods and mention the potential of applying a time deformation technique to account for genuine SV features of the process, as in Conley, Hansen, Luttmer, and Scheinkman [109]. In addition, the Aït-Sahalia [3,4] closed-form polynomial expansions for discretely-sampled diffusions are reviewed along with the Schaumburg [235] extension to a general class of Markov processes with Lévy-type generators. Meanwhile, Bibby, Jacobsen, and Sørensen [61] survey the extensive statistics literature on estimating functions for diffusion-type models and Bandi and Phillips [42] explicitly consider dealing with nonstationary processes (see also the work of Bandi [39], Bandi and Nguyen [41], and Bandi and Phillips [43,44]).

The characteristic function based inference technique has been particularly widely adopted due to the natural fit with the exponentially affine model class which provides essentially closed-form solutions for many pricing applications. Consequently, we dedicate a separate section to this approach.

Characteristic Functions Singleton [242] proposes to exploit the information contained in the conditional characteristic function of the state vector X,

$$\phi(iu, X(t), \Psi) = \mathrm{E}\left[\mathrm{e}^{iu' X(t+1)} \middle| X(t)\right] , \qquad (65)$$

to pursue maximum likelihood estimation of affine term structure models. In Equation (65) we highlight the dependence of the characteristic function on the unknown parameter vector Ψ. When X is an affine (jump-)diffusion process, ϕ has the exponential affine form,

$$\phi(iu, X(t), \Psi) = \mathrm{e}^{\alpha_t(u)+\beta_t(u)' X(t)} , \qquad (66)$$

where the functions α and β solve a system of ODEs. As such, the transition density f_X is known explicitly up to an inverse-Fourier transformation of the characteristic function (65),

$$f_X(X(t+1)\big|X(t); \Psi) = \frac{1}{\pi^N} \int_{\mathbb{R}^N_+} \mathrm{Re}\big[\mathrm{e}^{-iu' X(t+1)} \phi(iu, X(t), \Psi)\big] \mathrm{d}u \, . \qquad (67)$$

Singleton shows that Gauss–Legendre quadrature with a relatively small number of quadrature points allows to accurately evaluate the integral in Eq. (67) when X is univariate. As such, the method readily delivers efficient estimates of the parameter vector, Ψ, subject to an auxiliary assumption, namely that the state vector may be extracted by assuming that a pre-specified set of security prices is observed without error while the remainder have non-trivial error terms.

When X is multivariate the Fourier inversion in Eq. (67) is computationally more demanding. Thus, when estimating multi-dimensional systems Singleton suggests focusing on the conditional density function of the individual elements of X, but conditioned on the full state vector,

$$f_{X_j}(X_j(t+1)|X(t); \Psi) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-i\omega \mathbf{I}_j' X(t+1)} \phi(i\omega \mathbf{I}_j, X(t), \Psi) \mathrm{d}\omega \, , \qquad (68)$$

where the vector $\mathbf{I}_j$ has 1 in the jth element and zero elsewhere so that the jth element of X is $X_j(t+1) = \mathbf{I}_j' X(t+1)$. Maximization of the likelihood function obtained from f_{X_j}, for a fixed j, will often suffice to obtain a consistent estimate of Ψ. Exploiting more than one of the conditional densities (68) will result in more efficient Ψ estimate. For instance, the scores of multiple univariate log-likelihood functions, stacked in a vector, yield moment conditions that allow for generalized method of moment (GMM) estimation of the system. Alternatively, Joslin [192] proposes a change-of-measure transformation which reduces the oscillatory behavior of the integrand in Eq. (67). When using this transformation, Gauss-Hermite quadrature more readily provides a solution to the integral in (67) even if the state vector X is multi-dimensional, thus facilitating full ML estimation of the system.

Related, several studies have pursued GMM estimation of affine processes using characteristic functions. Definition (65) yields the moment condition

$$\mathrm{E}\left[(\phi(iu, X(t), \Psi) - \mathrm{e}^{iu' X(t+1)}) z(u, X(t))\right] = 0 \, , \qquad (69)$$

where X is an N-dimensional (jump-)diffusion, $u \in \mathbb{R}^N$, and z is an instrument function. When X is affine, the characteristic function takes the exponential form (66). Different choices of u and z yield a set of moment conditions that can be used for GMM estimation and inference. Singleton [242] derives the optimal instrument in terms of delivering efficient estimates. Carrasco et al. [86] approximate the optimal instrument with a set of basis functions that do not require the knowledge of the conditional likelihood function, thus avoiding one of the assumptions invoked by Singleton. Further, they build on Carrasco and Florens [87] to implement estimation using a continuum of moment conditions, which yields maximum-likelihood efficiency. Other applications of GMM-characteristic function methods to affine (jump-) diffusions for equity index returns are in Chacko and Viceira [88] and Jiang and Knight [184].

In some cases the lack of closed-form expressions for the moment condition in Eq. (69) can hinder GMM estimation. In these cases the expectation in Eq. (69) can be evaluated by Monte Carlo integration. This is accomplished by simulating a long sample from the discretized process for a given value of the coefficient vector Ψ. The parameter Ψ is then estimated via the simulated method of moments (SMM) of McFadden [206] and Duffie and Singleton [131]. Singleton [242] proposes SMM characteristic function estimators that exploit the special structure of affine term structure models.

Efficient Estimation of General Continuous-Time Processes

A number of recent approaches offer excellent flexibility in terms of avoiding approximations to the continuous-time model-implied transition density while still facilitating efficient estimation of the evolution of the latent state vector for the system.

Maximum Likelihood with Characteristic Functions

Bates [58] develops a filtration-based maximum likelihood estimation method for affine processes. His approach relies on Bayes' rule to recursively update the joint characteristic function of latent variables and data conditional on past data. He then obtains the transition density by Fourier inversion of the updated characteristic function.

Denote with $y(t)$ and $X(t)$ the time-t values of the observable variable and the state vector, respectively, and let $Y(t) \equiv \{y(1), \ldots, y(t)\}$ be the data observed up to time t. Consider the case in which the characteristic function of $z(t+1) \equiv (y(t+1), X(t+1))$ conditional on $z(t) \equiv (y(t), X(t))$, is an exponential affine function of $X(t)$:

$$\begin{aligned}\phi(is, iu, z(t), \Psi) &= \mathrm{E}\left[\mathrm{e}^{is'y(t+1)+iu'X(t+1)}\middle|z(t)\right]\\ &= \mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))'X(t)} \,. \qquad (70)\end{aligned}$$

Next, determine the value of the characteristic function conditional on the observed data $Y(t)$:

$$\begin{aligned}&\phi(is, iu, Y(t), \Psi)\\ &\quad= \mathrm{E}\left[\mathrm{E}\left[\mathrm{e}^{is'y(t+1)+iu'X(t+1)}\middle|z(t)\right]\middle|Y(t)\right]\\ &\quad= \mathrm{E}\left[\mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))'X(t)}\middle|Y(t)\right]\\ &\quad= \mathrm{e}^{\alpha(is,iu,y(t))}\psi(\beta(is, iu, y(t)), Y(t), \Psi) \,, \qquad (71)\end{aligned}$$

where $\psi(iu, Y(t), \Psi) \equiv \mathrm{E}\left[\mathrm{e}^{iu'X(t)}\middle|Y(t)\right]$ denotes the (marginal) characteristic function for the state vector conditional on the observed data. Fourier inversion then yields the conditional density for the observation $y(t+1)$ conditional on $Y(t)$:

$$\begin{aligned}&f_y(y(t+1)|Y(t);\Psi)\\ &\quad= \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi(is, 0, Y(t), \Psi)\mathrm{d}s \,. \qquad (72)\end{aligned}$$

The next step updates the characteristic function ψ (Bartlett [48]):

$$\begin{aligned}\psi(iu, Y(t+1), \Psi) &= \frac{1}{2\pi f_y(y(t+1)|Y(t);\Psi)}\\ &\quad\cdot \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi(is, iu, Y(t), \Psi)\mathrm{d}s \,. \qquad (73)\end{aligned}$$

To start the recursion, Bates initializes ψ at the unconditional characteristic function of the latent variable X. The log-likelihood function is then given by

$$\begin{aligned}\log \mathcal{L}(Y(T);\Psi) &= \log(f_y(y(1);\Psi)\\ &\quad+ \sum_{t=2}^{T} \log(f_y(y(t)|Y(t-1);\Psi)) \,. \qquad (74)\end{aligned}$$

A nice feature is that the method provides a natural solution to the filtering problem. The filtered estimate of the latent state X and its variance are computed from the first and second derivatives of the moment generating function $\psi(u, Y(t);\Psi)$ in Eq. (73), evaluated at $u = 0$:

$$\begin{aligned}\mathrm{E}[X(t+1)|Y(t+1);\Psi] &= \frac{1}{2\pi f_y(y(t+1)|Y(t);\Psi)}\\ &\quad\times \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi_u(is, 0, Y(t);\Psi)\mathrm{d}s \qquad (75)\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[X(t+1)|Y(t+1);\Psi] &= \frac{1}{2\pi f_y(y(t+1)|Y(t);\Psi)}\\ \times \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi_{uu}(is, 0, Y(t);\Psi)\mathrm{d}s&\\ &\quad- \{\mathrm{E}[X(t+1)|Y(t+1)]\}^2 \,. \qquad (76)\end{aligned}$$

A drawback is that at each step t of the iteration the method requires storage of the entire characteristic function $\psi(iu, Y(t);\Psi)$. To deal with this issue Bates recommends to approximate the true ψ with the characteristic function of a variable with a two-parameter distribution. The choice of the distribution depends on the X-dynamics while the two parameters of the distribution are determined by the conditional mean $\mathrm{E}[X(t+1)|Y(t+1);\Psi]$ and variance $\mathrm{Var}[X(t+1)|Y(t+1);\Psi]$ given in Equations (75)–(76).

In his application Bates finds that the method is successful in estimating different flavors of the SV jump-diffusion for a univariate series of daily 1953–1996 S&P 500 returns. In particular, he shows that the method obtains estimates that are equally, if not more, efficient compared to the efficient method of moments and Markov Chain Monte Carlo methods described below. Extensions of the method to multivariate processes are theoretically possible, but they require numerical integration of multi-dimensional functions, which is computationally demanding.

Simulated Maximum Likelihood In Sect. "Filtering the Latent State Variable Directly During Estimation" we discussed methods for simulated ML estimation and inference in discrete-time SV models. Pedersen [224] and Santa-Clara [234] independently develop a simulated maximum likelihood (SML) method to estimate continuous-time diffusion models. They divide each interval in between two consecutive data points X_{t+1} and X_t into M sub-intervals of length $\Delta = 1/M$ and they discretize the X process using the Euler scheme,

$$\begin{aligned} X_{t+(i+1)\Delta} = X_{t+i\Delta} &+ \mu(X_{t+i\Delta})\Delta \\ &+ \Sigma(X_{t+i\Delta})\sqrt{\Delta}\varepsilon_{t+(i+1)\Delta} \,, \\ &\quad i = 0, \dots, M-1 \,, \end{aligned} \tag{77}$$

where μ and Σ are the drift and diffusion terms of the X process and ε is multivariate normal with mean zero and identity variance matrix. The transition density of the discretized process is multivariate normal with mean μ and variance matrix $\Sigma\Sigma'$. As Δ goes to zero, this density converges to that of the continuous-time process X. As such, the transition density from X_t to X_{t+1} is given by

$$\begin{aligned} f_X(X_{t+1}|X_t;\Psi) = \int & f_X(X_{t+1}|X_{t+1-\Delta};\Psi) \\ & \times f_X(X_{t+1-\Delta}|X_t;\Psi)\mathrm{d}X_{t+1-\Delta} \,. \end{aligned} \tag{78}$$

For sufficiently small values of Δ the first term in the integrand, $f_X(X_{t+1}|X_{t+1-\Delta};\Psi)$, is approximated by the transition density of the discretized process, while the second term, $f_X(X_{t+1-\Delta}|X_t;\Psi)$, is a multi-step-ahead transition density that can be computed from the recursion from X_t to $X_{t+1-\Delta}$. Writing the right-hand side of Eq. (78) as a conditional expectation yields

$$f_X(X_{t+1}|X_t;\Psi) = E_{X_{t+1-\Delta}|X_t}\left[f_X(X_{t+1}|X_{t+1-\Delta};\Psi)\right] . \tag{79}$$

The expectation in Eq. (79) can be computed by Monte Carlo integration over a large number of paths for the process X, simulated via the Euler scheme (77). As Δ vanishes, the Euler scheme is consistent. Thus, when the size of the simulated sample increases the sample average of the function f_X, evaluated at the random draws of $X_{t+1-\Delta}$, converges to the true transition density. Application of the principles in Bladt and Sørensen [64] may well be useful in enhancing the efficiency of the simulation scheme and hence the actual efficiency of the inference procedure in practice.

Brandt and Santa-Clara [75] apply the SML method to estimate a continuous-time model of the joint dynamics of interest rates in two countries and the exchange rate between the two currencies. Piazzesi [227] extends the SML approach for jump-diffusion processes with time-varying jump intensity. She considers a high-frequency policy rule based on yield curve information and an arbitrage-free bond market and estimates the model using 1994–1998 data on the Federal Reserve target rate, the six-month LIBOR rate, and swap yields.

An important issue is how to initialize any unobserved component of the state vector, $X(t)$, such as the volatility state at each observation to provide a starting point for the next Monte Carlo integration step. This may be remedied through application of the particle filter, as mentioned earlier and discussed below in connection with MCMC estimation. Another possibility is, as also indicated previously, to extract the state variable through inversion from derivatives prices or yields assumed observed without pricing errors.

Indirect Inference There are also other method-of-moments strategies to estimate finitely-sampled continuous-time processes of a general type. One prominent approach approximates the unknown transition density for the continuous-time process with the density of a semi-nonparametric (SNP) auxiliary model. Then one can use the score function of the auxiliary model to form moment conditions for the parameter vector Ψ of the continuous-time model. This approach yields the efficient method of moments estimator (EMM) of Gallant and Tauchen [154], Gallant et al. [150], and Gallant and Long [152], and the indirect inference estimator of Gouriéroux et al. [162] and Smith [243].

To fix ideas, suppose that the conditional density for a continuous-time return process r (the 'structural' model) is unknown. We intend to approximate the unknown density with a discrete-time model (the 'auxiliary' model) that is tractable and yet sufficiently flexible to accommodate the systematic features of the actual data sample well. A parsimonious auxiliary density for r embeds ARMA and EGARCH leading terms to capture the conditional mean and variance dynamics. There may be residual excess skewness and kurtosis that elude the ARMA and EGARCH forms. As such, the auxiliary density is rescaled using a nonparametric polynomial expansion of order K, which yields

$$g_K(r(t)|x(t);\xi) = \left(\nu + (1-\nu) \times \frac{[P_K(z(t),x(t))]^2}{\int_{\mathbb{R}}[P_K(z(t),x(t))]^2\phi(u)\mathrm{d}u}\right)\frac{\phi(z(t))}{\sqrt{h(t)}} , \tag{80}$$

where ν is a small constant, $\phi(.)$ is the standard normal density, $x(t)$ contains lagged return observations, and

$$z(t) = \frac{r(t)-\mu(t)}{\sqrt{h(t)}} , \tag{81}$$

$$\mu(t) = \phi_0 + ch(t) + \sum_{i=1}^{s}\phi_i r(t-1) + \sum_{i=1}^{u}\delta_i \varepsilon(t-1) , \tag{82}$$

$$\log h(t) = \omega + \sum_{i=1}^{p}\beta_i \log h(t-1) + (1+\alpha_1 L + \cdots + \alpha_q L^q) \times \left[\theta_1 z(t-1) + \theta_2(b(z(t-1)) - \sqrt{2/\pi})\right] , \tag{83}$$

$$P_K(z,x) = \sum_{i=0}^{K_z} a_i(x)z^i = \sum_{i=0}^{K_z}\left(\sum_{|j|=0}^{K_x} a_{ij}x^j\right) z^i , \quad a_{00} = 1 . \tag{84}$$

Here j is a multi-index vector, $x^j \equiv (x_1^{j_1},\dots,x_M^{j_M})$, and $|j| \equiv \sum_{m=1}^{M} j_m$. The term $b(z)$ is a smooth (twice-differentiable) function that closely approximates the absolute value operator in the EGARCH variance equation.

In practice, the representation of P_K is given by Hermite orthogonal polynomials. When the order K of the expansion increases, the auxiliary density will approximate the data arbitrarily well. If the structural model is indeed the true data generating process, then the auxiliary density will converge to that of the structural model. For a given K, the QML estimator $\hat{\xi}$ for the auxiliary model coefficient satisfies the score condition

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial \xi} = 0 . \tag{85}$$

Suppose now that the structural model is correct and Ψ_0 is the true value of its coefficient vector. Consider a series $\{r(t;\Psi), x(t;\Psi)\}$, $t = 1,\dots,\mathcal{T}(T)$, simulated from the structural model. Then we expect that the score condition (85) holds when evaluated by averaging over the simulated returns rather than over the actual data:

$$m_{\mathcal{T}(T)}(\Psi_0,\hat{\xi}) = \frac{1}{\mathcal{T}(T)}\sum_{t=1}^{\mathcal{T}(T)}\frac{\partial \log g_K(r(t,\Psi_0)|x(t,\Psi_0);\hat{\xi})}{\partial \xi} \approx 0 . \tag{86}$$

When T and $\mathcal{T}(T)$ tend to infinity, condition (86) holds exactly.

Gallant and Tauchen [154] propose the EMM estimator $\hat{\Psi}$ defined via

$$\hat{\Psi} = \arg\min_{\Psi} m_{\mathcal{T}(T)}(\Psi,\hat{\xi})'\,\hat{W}_T m_{\mathcal{T}(T)}(\Psi,\hat{\xi}) , \tag{87}$$

where the weighting matrix $\hat{W}_T$ is a consistent estimate of the inverse asymptotic covariance matrix of the auxiliary score function, e. g., the inverse outer product of the SNP gradient:

$$\hat{W}_T^{-1} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial \xi}\right] \times \left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial \xi}\right]' . \tag{88}$$

An important advantages of the technique is that EMM estimates achieve the same degree of efficiency as the ML procedure, when the score of the auxiliary model asymptotically spans the score of the true model. It also delivers powerful specification diagnostics that provide guidance in the model selection. Gallant and Tauchen [154] show that the EMM estimator is asymptotically normal. Further, under the assumption that the structural model is correctly specified, they derive a χ^2 statistic for the test of over-identifying restrictions. Gallant et al. [150] normalize the vector $m_{\mathcal{T}(T)}(\hat{\Psi},\hat{\xi})$ by its standard error to obtain a vector of score t-ratios. The significance of the individual score elements is often informative of the source of model mis-specification, with the usual caveat that failure to capture one characteristic of the data may result in the significance of a moment condition that pertains to a coefficient not directly related to that characteristic (due to correlation in the moment conditions). Finally, EMM provides a straightforward solution to the problem of filtering and forecasting the latent return variance process V,

i. e., determining the conditional densities $f(V(t)|x(t),\Psi)$ and $f(V(t+j)|x(t),\Psi)$, $j \geq 0$. This is accomplished through the *reprojection* method discussed in, e. g., Gallant and Long [152] and Gallant and Tauchen [155]. In applications to dynamic term structure models, the same method yields filtered and forecasted values for the latent state variables.

The reprojection method assumes that the coefficient vector Ψ is known. In practice, Ψ is fixed at the EMM estimate $\hat{\Psi}$. Then one simulates a sample of returns and latent variables from the structural model and fits the auxiliary model on the simulated data. This is equivalent to the first step of the EMM procedure except that, in the reprojection step, we fit the auxiliary model assuming the structural model is correct, rather than using actual data. The conditional density of the auxiliary model, estimated under the null, approximates the unknown density of the structural model:

$$g_K(r(t+j)|x(t);\tilde{\xi}) \approx f(r(t+j)|x(t);\hat{\Psi}), \quad j \geq 0\,, \tag{89}$$

where $\tilde{\xi}$ is the QML estimate of the auxiliary model coefficients obtained by fitting the model on simulated data. This approach yields filtered estimates and forecasts for the conditional mean and variance of the return via

$$\mathrm{E}\left[r(t+j)|x(t);\hat{\Psi}\right] = \int y g_K(y|x(t);\tilde{\xi})\mathrm{d}y\,, \tag{90}$$

$$\mathrm{Var}\left[r(t+j)|x(t);\hat{\Psi}\right] = \int \left(y - \mathrm{E}\left[r(t+j)|x(t);\hat{\Psi}\right]\right)^2 \times g_K(y|x(t);\tilde{\xi})\mathrm{d}y\,. \tag{91}$$

An alternative approach consists in fitting an auxiliary model for the latent variable (e. g., the return conditional variance) as a function of current and lagged returns. It is straightforward to estimate such model using data on the latent variable and the associated returns simulated from the structural model with the EMM coefficient $\hat{\Psi}$. Also in this case the auxiliary model density approximates the true one, i. e.,

$$g_K^V(V(t+j)|x(t);\tilde{\xi}) \approx f^V(V(t+j)|x(t);\hat{\psi})\,, \quad j \geq 0. \tag{92}$$

This approach yields a forecast for the conditional variance process,

$$\mathrm{E}\left[V(t+j)|x(t);\hat{\Psi}\right] = \int v g_K^V(v|x(t);\tilde{\xi})\mathrm{d}v\,. \tag{93}$$

In sum, reprojection is a simulation approach to implement a non-linear Kalman-filter-type technique, which yields effective forecasts for the unobservable state vector.

The indirect inference estimator by Gouriéroux et al. [162] and Smith [243] is closely related to the EMM estimator. Indirect inference exploits that the following two quantities should be close when the structural model is correct and the data are simulated at the true parameter Ψ_0: (i) the QML estimator $\hat{\xi}$ for the auxiliary model computed from actual data; (ii) the QML estimator $\hat{\xi}(\Psi)$ for the auxiliary model fitted on simulations from the structural model. Minimizing the distance between $\hat{\xi}$ and $\hat{\xi}(\Psi)$ in an appropriate metric yields the indirect inference estimator for Ψ. Similar to EMM, asymptotic normality holds and a χ^2 test for over-identifying restrictions is available. However, the indirect inference approach is computationally more demanding, because finding the value of Ψ that minimizes the distance function requires re-estimating the auxiliary model on a different simulated sample for each iteration of the optimization routine. EMM does not have this drawback, since the EMM objective function is evaluated at the same fitted score at each iteration. Nonetheless, there may well be circumstances where particular auxiliary models are of primary economic interest and estimation based on the corresponding moment conditions may serve as a useful diagnostic tool for model performance in such directions.

Several studies have used EMM to fit continuous-time SV jump-diffusion models for equity index returns, e. g., Andersen et al. [15], Benzoni [59], Chernov and Ghysels [93], and Chernov et al. [94,95]. Andersen and Lund [28] and Andersen et al. [16] use EMM to estimate SV jump-diffusion models for the short-term interest rate. Ahn et al. [1,2], Brandt and Chapman [71], and Dai and Singleton [114] fit different flavors of multi-factor dynamic term structure models. Andersen et al. [27] document the small-sample properties of the efficient method of moments estimator for stationary processes, while Duffee and Stanton [128] study its properties for near unit-root processes. A. Ronald Gallant and George E. Tauchen at Duke University have prepared well-documented general-purpose EMM and SNP packages, available for download at the web address ftp.econ.duke.edu in the directories pub/get/emm and pub/get/snp. In applications it is often useful to customize the SNP density to allow for a more parsimonious fit of the data under investigation. For instance, Andersen et al. [15,16], Andersen and Lund [28], and Benzoni [59] rely on the SNP density (80)–(84).

Markov Chain Monte Carlo The MCMC method provides a Bayesian solution to the inference problem for a dynamic asset pricing model. The approach treats the model coefficient Ψ as well as the vector of latent state variables X as random variables and computes the posterior distribution $f(\Psi, X|Y)$, conditional on certain observable variables Y, predicted by the model. The setting is sufficiently general to deal with a wide range of situations. For instance, X and Y can be the (latent) volatility and (observable) return processes as is the case of an SV model for asset returns. Or X and Y can be the latent state vector and observable yields in a dynamic term structure model.

The posterior distribution $f(\Psi, X|Y)$ is the main tool to draw inference not only on the coefficient Ψ but also on the latent vector X. Since $f(\Psi, X|Y)$ is unknown in closed-form in relevant applications, MCMC relies on a simulation (a Markov Chain) from the conditional density $f(\Psi, X|Y)$ to compute mode, mean, and standard deviations for the model coefficients and state variables via the Monte Carlo method.

The posterior $f(\Psi, X|Y)$ is analytically untractable and extremely high-dimensional, so that simulation directly from $f(\Psi, X|Y)$ is typically infeasible. The MCMC approach hinges on the Clifford–Hammersley theorem, which determines conditions under which the posterior $f(\Psi, X|Y)$ is uniquely determined by the marginal posterior distributions $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. In turn, the posteriors $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$ are determined by a set or univariate posterior distributions. Specifically, denote with $\Psi(i)$ the ith element of the coefficient Ψ, $i = 1, \ldots, K$, and with $\Psi(-i)$ the vector consisting of all elements in Ψ except for the ith one. Similarly denote with $X(t)$ the tth row of the state vector, $t = 1, \ldots, T$, and with $X(-t)$ the rest of the vector. Then the Clifford–Hammersley theorem allows to characterize the posterior $f(\Psi, X|Y)$ via $K + T$ univariate posteriors,

$$f(\Psi(i)|\Psi(-i), X, Y)\ , \quad i = 1, \ldots, K \tag{94}$$

$$f(X(t)|X(-t), \Psi, Y)\ , \quad t = 1, \ldots, T\,. \tag{95}$$

The construction of the Markov Chain relies on the so-called Gibbs sampler. The first step of the algorithm consists in choosing initial values for the coefficient and the state, Ψ_0 and X_0. When (one of or both) the multi-dimensional posteriors are tractable, the Gibbs sampler generates values Ψ_1 and X_1 directly from $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. Alternatively, each element of Ψ_1 and X_1 is drawn from the univariate posteriors (94)–(95). Some of these posteriors may also be analytically intractable or efficient algorithms to draw from these posteriors may not exist. In such cases the Metropolis-Hastings algorithm ensures that the simulated sample is consistent with the posterior target distribution. Metropolis-Hastings sampling consists of an accept-reject procedure of the draws from a ‘proposal’ or ‘candidate’ tractable density, which is used to approximate the unknown posterior (see, e. g., Johannes and Polson [187]).

Subsequent iterations of Gibbs sampling, possibly in combination with the Metropolis-Hastings sampling, yield a series of ‘sweeps’ $\{\Psi_s, X_s\}$, $s = 1, \ldots, S$, with limiting distribution $f(\Psi, X|Y)$. A long number of sweeps may be necessary to ‘span’ the whole posterior distribution and obtain convergence due to the serial dependence of subsequent draws of coefficients and state variables. When the algorithm has converged, additional simulations provide a sample from the joint posterior distribution.

The MCMC approach has several advantages. First, the inference automatically accounts for parameter uncertainty. Further, the Markov Chain provides a direct and elegant solution to the *smoothing* problem, i. e., the problem of determining the posterior distribution for the state vector X conditional on the entire data sample, $f(X(t)|Y(1), \ldots, Y(T), \Psi)$, $t = 1, \ldots, T$. The limitation on the approach is largely that efficient sampling schemes for the posterior distribution must be constructed for each specific problem at hand which by nature is case specific and potentially cumbersome or inefficient. Nonetheless, following the development of more general simulation algorithms, the method has proven flexible for efficient estimation of a broad class of important models.

One drawback is that MCMC does not deliver an immediate solution to the *filtering* problem, i. e., determining $f(X(t)|Y(1), \ldots, Y(t), \Psi)$, and the *forecasting* problem, i. e., determining $f(X(t+j)|Y(1), \ldots, Y(t), \Psi)$, $j > 0$. However, recent research is overcoming this limitation through the use of the ‘particle filter’. Bayes rule implies

$$\begin{aligned} f(X(t+1)|Y(1), \ldots, Y(t+1), \Psi) \propto f(Y(t+1)|\\ X(t+1), \Psi) f(X(t+1)|Y(1), \ldots, Y(t), \Psi)\ , \end{aligned} \tag{96}$$

where the symbol $\propto$ denotes ‘proportional to’. The first density on the right-hand side of Eq. (96) is determined by the SV model and it is often known in closed form. In contrast, the second density at the far-right end of the equation is given by an integral that involves the unknown filtering density at the prior period, $f(X(t)|Y(1), \ldots, Y(t), \Psi)$:

$$\begin{aligned} f(X(t{+}1)|Y(1), \ldots, Y(t), \Psi) = \int f(X(t{+}1)|X(t), \Psi)\\ \times f(X(t)|Y(1), \ldots, Y(t), \Psi) \mathrm{d}X(t)\,. \end{aligned} \tag{97}$$

The particle method relies on simulations to construct a finite set of weights $w^i(t)$ and particles $X^i(t)$, $i = 1, \ldots, N$, that approximate the unknown density with a finite sum,

$$f(X(t)|Y(1),\dots,Y(t),\Psi) \approx \sum_{i=1}^{N} w^i(t)\delta_{X^i(t)}\,, \tag{98}$$

where the Dirac function $\delta_{X^i(t)}$ assigns mass one to the particle $X^i(t)$. Once the set of weights and particles are determined, it is possible to re-sample from the discretized distribution. This step yields a simulated sample $\{X^s(t)\}_{s=1}^S$ which can be used to evaluate the density in Eq. (97) via Monte Carlo integration:

$$\begin{aligned} &f(X(t+1)|Y(1),\dots,Y(t),\Psi) \\ &\qquad \approx \frac{1}{S}\sum_{s=1}^{S} f(X(t+1)|X^s(t),\Psi)\,. \end{aligned} \tag{99}$$

Equation (99) solves the forecasting problem while combining formulas (96) and (99) solves the filtering problem. The challenge in practical application of the particle filter is to identify an accurate and efficient algorithm to construct the set of particles and weights. We point the interested reader to Kim et al. [193], Pitt and Shephard [228] and Johannes and Polson [187] for a discussion on how to approach this problem.

The usefulness of the MCMC method to solve the inference problem for SV models has been evident since the early work by Jacquier et al. [180], who develop an MCMC algorithm for the logarithmic SV model. Jacquier et al. [181] provide extensions to correlated and non-normal error distributions. Kim et al. [193], Pitt and Shephard [228] and Chib et al. [96] develop simulation-based methods to solve the filtering problem, while Chib et al. [97] use the MCMC approach to estimate a multivariate SV model. Elerian et al. [135] and Eraker [140] discuss how to extend the MCMC inference method to a continuous-time setting. Eraker [140] uses the MCMC approach to estimate an SV diffusion process for interest rates, while Jones [189] estimates a continuous-time model for the spot rate with non-linear drift function. Eraker et al. [142] estimate an SV jump-diffusion process using data on S&P 500 return while Eraker [141] estimates a similar model using joint data on options and underlying S&P 500 returns. Li et al. [196] allow for Lévy-type jumps in their model. Collin-Dufresne et al. [104] use the MCMC approach to estimate multi-factor affine dynamic term structure model using swap rates data. Johannes and Polson [186] give a comprehensive survey of the still ongoing research on the use of the MCMC approach in the general nonlinear jump-diffusion SV setting.

Estimation from Option Data

Options' payoffs are non-linear functions of the underlying security price. This feature renders options highly sensitive to jumps in the underlying price and to return volatility, which makes option data particularly useful to identify return dynamics. As such, several studies have taken advantage of the information contained in option prices, possibly in combination with underlying return data, to estimate SV models with or without discontinuities in returns and volatility.

Applications to derivatives data typically require a model for the pricing errors. A common approach is to posit that the market price of an option, O^*, normalized by the underlying observed security price S^*, is the sum of the normalized model-implied option price, O/S^*, and a disturbance term ε (e. g., Renault [230]):

$$\frac{O^*}{S^*} = \frac{O(S^*,V,K,\tau,\Psi)}{S^*} + \varepsilon\,, \tag{100}$$

where V is the latent volatility state, K is the option strike price, τ is time to maturity, and Ψ is the vector with the model coefficients. A pricing error ε could arise for several reasons, including measurement error (e. g., price discreteness), asynchroneity between the derivatives and underlying price observations, microstructure effects, and perhaps most importantly specification error. The structure imposed on ε depends on the choice of a specific 'loss function' used for estimation (e. g., Christoffersen and Jacobs [98]). Several studies have estimated the coefficient vector Ψ by minimizing the sum of the squared option pricing errors normalized by the underlying price S^*, as in Eq. (100). Others have focused on either squared dollar pricing errors, or squared errors normalized by the options market price (instead of S^*). The latter approach has the advantage that a $1 error on an expensive in-the-money option carries less weight than the same error on a cheaper out-of-the-money contract. The drawback is that giving a lot of weight to the pricing errors on short-maturity deep-out-of-the-money options could bias the estimation results. Finally, the common practice of expressing option prices in terms of their Black-Scholes implied volatilities has inspired other scholars to minimize the deviations between Black-Scholes implied volatilities inferred from model and market prices (e. g., Mizrach [216]). An alternative course is to form a moment-based loss function and follow a GMM- or SMM-type approach to estimate Ψ. To this end moment conditions stem from distributional assumptions on the pricing error ε (e. g., $\mathrm{E}[\varepsilon] = 0$) or from the scores of a reduced-form model that approximates the data.

In estimating the model, some researchers have opted to use a panel of options consisting of contracts with multiple strikes and maturities across dates in the sample period. This choice brings a wealth of information on the cross-sectional and term-structure properties of the implied volatility smirk into the analysis. Others rely on only one option price observation per time period, which shifts the focus to the time-series dimension of the data. Some studies re-estimate the model on a daily basis rather than seeking a single point estimate for the coefficient Ψ across the entire sample period. This ad hoc approach produces smaller in-sample pricing errors, which can be useful to practitioners, but at the cost of concealing specification flaws by over-fitting the model, which tends to hurt out-of-sample performance. The different approaches are in part dictated by the intended use of the estimated system as practitioners often are concerned with market making and short-term hedging while academics tend to value stable relations that may form the basis for consistent modeling of the dominant features of the system over time.

Early contributions focus on loss functions based on the sum of squared option pricing errors and rely entirely on option data for estimation. This approach typically yields an estimate of the model coefficient Ψ that embeds an adjustment for risk, i. e., return and volatility dynamics are identified under the risk-neutral rather than the physical probability measure. For instance, Bates [56] considers an SV jump-diffusion model for Deutsche Mark foreign currency options and estimates its coefficient vector Ψ via nonlinear generalized least squares of the normalized pricing errors with daily option data from January 1984 to June 1991. A similar approach is followed by Bates [57] who fits an SV model with two latent volatility factors and jumps using daily data on options on the S&P 500 futures from January 1988 to December 1993. Bakshi et al. [34] focus on the pricing and hedging of daily S&P 500 index options from June 1988 to May 1991. In their application they re-calibrate the model on a daily basis by minimizing the sum of the squared dollar pricing errors across options with different maturities and strikes. Huang and Wu [173] explore the pricing implications of the time-changed Lévy process by Carr and Wu [84] for daily S&P 500 index options from April 1999 to May 2000. Their Lévy return process allows for discontinuities that exhibit higher jump frequencies compared to the finite-intensity Poisson jump processes in Equations (37)–(41). Further, their model allows for a random time change, i. e., a monotonic transformation of the time variable which generates SV in the diffusion and jump components of returns. In contrast, Bakshi et al. [35] fit an SV jump-diffusion model by SMM using daily data on long-maturity S&P 500 options (LEAPS).

More recent studies have relied on joint data on S&P 500 option prices and underlying index returns, spanning different periods, to estimate the model. This approach forces the same model to price securities in two different markets and relies on information from the derivatives and underlying securities to better pin down model coefficients and risk premia. For instance, Eraker [141] and Jones [190] fit different flavors of the SV model (with and without jumps, respectively) by MCMC. Pan [220] follows a GMM approach to estimate an SV jump-diffusion model using weekly data. She relies on a single at-the-money option price observation each week, which identifies the level of the latent volatility state variable (i. e., at each date she fixes the error term ε at zero and solves Eq. (100) for V). Aït-Sahalia and Kimmel [7] apply Aït-Sahalia's [4] method to approximate the likelihood function for a joint sample of options and underlying prices. Chernov and Ghysels [93] and Benzoni [59] obtain moment conditions from the scores of a SNP auxiliary model. Similarly, other recent studies have found it useful to use joint derivatives and interest rate data to fit dynamic term structure models, e. g., Almeida et al. [9], and Bikbov and Chernov [62].

Finally, a different literature has studied the option pricing implications of a model in which asset return volatility is a deterministic function of the asset price and time, e. g., Derman and Kani [119], Dupire [134], Rubinstein [233], and Jackwerth and Rubinstein [177]. Since volatility is not stochastic in this setting, we do not review these models here and point the interested reader to, e. g., [133] for an empirical analysis of their performance.

Future Directions

In spite of much progress in our understanding of volatility new challenges lie ahead. In recent years a wide array of volatility-sensitive products has been introduced. The market for these derivatives has rapidly grown in size and complexity. Research faces the challenge to price and hedge these new products. Moreover, the recent developments in model-free volatility modeling have effectively given empirical content to the latent volatility variable, which opens the way for a new class of estimation methods and specification tests for SV systems. Related, improved volatility measures enable us to shed new light on the properties and implications of the volatility risk premium. Finally, more work is needed to better understand the linkage between fluctuations in economic fundamentals and low- and high-frequency volatility movements. We conclude this chapter by briefly reviewing some open issues in these four areas of research.

Volatility and Financial Markets Innovation

Volatility is a fundamental input to any financial and real investment decision. Markets have responded to investors' needs by offering an array of volatility-linked instruments. In 1993 the Chicago Board Option Exchange (CBOE) has introduced the VIX index, which measures the market expectations of near-term volatility conveyed by equity-index options. The index was originally computed using the Black-Scholes implied volatilities of eight different S&P 100 option (OEX) series so that, at any given time, it represented the implied volatility of a hypothetical at-the-money OEX option with exactly 30 days to expiration (see [257]). On September 22, 2003, the CBOE began disseminating price level information using a revised 'model-free' method for the VIX index. The new VIX is given by the price of a portfolio of S&P 500 index options and incorporates information from the volatility smirk by using a wider range of strike prices rather than just at-the-money series (see [77]). On March 26, 2004, trading in futures on the VIX Index started on the CBOE Futures Exchange (CFE) while on February 24, 2006, options on the VIX began trading on the Chicago Board Options Exchange. These developments have opened the way for investors to trade on option-implied measures of market volatility. The popularity of the VIX prompted the CBOE to introduce similar indices for other markets, e. g., the VXN NASDAQ 100 Volatility Index.

Along the way, a new over-the-counter market for volatility derivatives has also rapidly grown in size and liquidity. Volatility derivatives are contracts whose payments are expressed as functions of realized variance. Popular examples are variance swaps, which at maturity pay the difference between realized variance and a fixed strike price. According to estimates by BNP Paribas reported by the Risk [176] magazine, the daily trading volume for variance swaps on indices reached $4–5 million in vega notional (measured in dollars per volatility point) in 2006, which corresponds to payments in excess of $1 billion per percentage point of volatility on an annual basis (Carr and Lee [82]). Using variance swaps hedge fund managers and proprietary traders can easily place huge bets on market volatility.

Finally, in recent years credit derivatives markets have evolved in complexity and grown in size. Among the most popular credit derivatives are the credit default swaps (CDS), which provide insurance against the risk of default by a particular company. The buyer of a single-name CDS acquires the right to sell bonds issued by the company at face value when a credit event occurs. Multiple-name contracts can be purchased simultaneously through credit indices. For instance, the CDX indices track the credit spreads for different portfolios of North American companies while the iTraxx Europe indices track the spreads for portfolios of European companies. At the end of 2006 the notional amount of outstanding over-the-counter single- and multi-name CDS contracts stood at $19 and $10 trillion, respectively, according to the September 2007 Bank for International Settlements Quarterly Review.

These market developments have raised new interesting issues for research to tackle. The VIX computations based on the new model-free definition of implied volatility used by the CBOE requires the use of options with strike prices that cover the entire support of the return distribution. In practice, liquid options satisfying this requirement often do not exist and the CBOE implementation introduces random noise and systematic error into the index (Jiang and Tian [185]). Related, the VIX implementation entails a truncation, i. e., the CBOE discards illiquid option prices with strikes lying in the tails of the return distribution. As such, the notion of the VIX is more directly linked to that of corridor volatility [26]. In sum, robust implementation of a model free measure of implied volatility is still an open area of research. Future developments in this direction will also have important repercussions on the hedging practices for implied-volatility derivatives.

Pricing and hedging of variance derivatives is another active area of research. Variance swaps admit a simple replication strategy via static positions in call and put options on the underlying asset, similar to model-free implied volatility measures (e. g., [77,83]). In contrast, it is still an open area of research to determine the replication strategy for derivatives whose payoffs are non-linear function of realized variance, e. g., volatility swaps, which pay the square-root of realized variance, or call and put options on realized variance. [82] is an interesting paper in this direction.

Limited liability gives shareholders the option to default on the firm's debt obligation. As such, a debt claim has features similar to a short position in a put option. The pricing of corporate debt is therefore sensitive to the volatility of the firms' assets: higher volatility increases the probability of default and therefore reduces the price of debt and increases credit spreads. The insights and techniques developed in the SV literature could prove useful in credit risk modeling and applications (e. g., [179,248,260]).

The Use of Realized Volatility for Estimation of SV Models

Another promising line of research aims at extracting the information in RV measures for the estimation of dynamic asset pricing models. Early work along these lines includes Barndorff-Nielson and Shephard [51], who decompose RV into actual volatility and realized volatility error. They consider a state-space representation for the decomposition and apply the Kalmann filter to estimate different flavors of the SV model. Moreover, Bollerslev and Zhou [68] and Garcia et al. [156], build on the insights of Meddahi [210] to estimate SV diffusion models using conditional moments of integrated volatility. More recently, Todorov [252] generalizes the analysis for the presence of jumps.

Related, recent studies have started to use RV measures to test the implications of models previously estimated with lower-frequency data. Since RV gives empirical content to the latent quadratic variation process, this approach allows for a direct test of the model-implied restrictions on the latent volatility factor. Recent work along these lines includes Andersen and Benzoni [12], who use model-free RV measures to show that the volatility spanning condition embedded in some affine term structure models is violated in the US Treasury market. Christoffersen et al. [100] note that the Heston square-root SV model implies that the dynamics for the standard deviation process are conditionally Gaussian. They reject this condition by examining the distribution of the changes in the square-root RV measure for S&P 500 returns.

Volatility Risk Premium

More work is needed to better understand the link between asset return volatility and model risk premia. Also in this case, RV measures are a fruitful source of information to shed new light on the issue. Among the recent studies that pursue this venue is Bollerslev et al. [66], who exploit the moments of RV and option-implied volatility to gauge a measure of the volatility risk premium. Todorov [251] explores the variance risk premium dynamics using high-frequency S&P 500 index futures data and data on the VIX index. He finds the variance risk premium to vary significantly over time and to increase during periods of high volatility and immediately after big jumps in underlying returns. Carr and Wu [85] provide a broader analysis of the variance risk premium for five equity indices and 35 individual stocks. They find the premium to be large and negative for the indices while it is much smaller for the individual stocks. Further, they also find the premium to increase (in absolute value) with the level of volatility. Additional work on the volatility risk premium embedded in individual stock options is in Bakshi and Kapadia [36], Driessen et al. [123], and Duarte and Jones [126]. Other studies have examined the linkage between volatility risk premia and equity returns (e. g., [69]) and hedge-fund performance (e. g., [70]). New research is also examining the pricing of aggregate volatility risk in the cross-section of stock returns. For instance, Ang et al. [30] find that average returns are lower on stocks that have high sensitivities to innovations in aggregate volatility and high idiosyncratic volatility (see also the related work by Chen [90] Ang et al. [32], Bandi et al, Guo et al. [42]). This evidence is consistent with the findings of the empirical option pricing literature, which suggests that there is a negative risk premium for volatility risk. Intuitively, periods of high market volatility are associated to worsened investment opportunities and tend to coincide with negative stock market returns (the so-called leverage effect). As such, investors are willing to pay higher prices (i. e., accept lower expected returns) to hold stocks that do well in high-volatility conditions.

Determinants of Volatility

Finally, an important area of future research concerns the linkage between asset return volatility and economic uncertainty. Recent studies have proposed general equilibrium models that produce low-frequency fluctuations in conditional volatility, e. g., Campbell and Cochrane [80], Bansal and Yaron [47], McQueen and Vorkink [207], and Tauchen [246]. Related, Engle and Rangel [139] and Engle et al. [138] link macroeconomic variables and long-run volatility movements. It is still an open issue, however, to determine the process through which news about economic fundamentals are embedded into prices to generate high-frequency volatility fluctuations. Early research by Schwert [236] and Shiller [241] has concluded that the amplitude of the fluctuations in aggregate stock volatility is difficult to explain using simple models of stock valuation. Further, Schwert [236] notes that while aggregate leverage is significantly correlated with volatility, it explains a relatively small part of the movements in stock volatility. Moreover, he finds little evidence that macroeconomic volatility (measured by inflation and industrial production volatility) helps predict future asset return volatility. Model-free realized volatility measures are a useful tool to further investigate this issue. Recent work in this direction includes Andersen et al. [22] and Andersen and Bollerslev [17], who explore the linkage between news arrivals and exchange rates volatility, and Andersen and Benzoni [13], who investigate the determinants of bond yields volatility in the US Treasury market. Related, Balduzzi et al. [38] and Fleming and Remolona [146] study the reaction of trading volume, bid-ask spread, and price volatility to macroeconomic surprises in the US Treasury market, while Brandt and Kavajecz [74] and Pasquariello and Vega [222] focus instead on the price discovery process and explore the implications of order flow imbalances (excess buying or selling pressure) on day-to-day variation in yields.

Acknowledgments

We are grateful to Olena Chyruk, Bruce Mizrach (the Section Editor) and Neil Shephard for helpful comments and suggestions. Of course, all errors remain our sole responsibility. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. The work of Andersen is supported by a grant from the NSF to the NBER and support from CREATES funded by the Danish National Research Foundation.

Bibliography

Primary Literature

1. Ahn DH, Dittmar RF, Gallant AR (2002) Quadratic Term Structure Models: Theory and Evidence. Rev Finance Stud 15:243–288
2. Ahn DH, Dittmar RF, Gallant AR, Gao B (2003) Purebred or hybrid?: Reproducing the volatility in term structure dynamics. J Econometrics 116:147–180
3. Aït-Sahalia Y (2002) Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach. Econometrica 70:223–262
4. Aït-Sahalia Y (2007) Closed-Form Likelihood Expansions for Multivariate Diffusions. Annals of Statistics, forthcoming
5. Aït-Sahalia Y, Hansen LP, Scheinkman J (2004) Operator Methods for Continuous-Time Markov Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
6. Aït-Sahalia Y, Jacod J (2007) Testing for Jumps in a Discretely Observed Process. Annals of Statistics, forthcoming
7. Aït-Sahalia Y, Kimmel R (2007) Maximum Likelihood Estimation of Stochastic Volatility Models. J Finance Econ 83:413–452
8. Alizadeh S, Brandt MW, Diebold FX (2002) Range-Based Estimation of Stochastic Volatility Models. J Finance 57:1047–1091
9. Almeida CIR, Graveline JJ, Joslin S (2006) Do Options Contain Information About Excess Bond Returns? Working Paper, UMN, Fundação Getulio Vargas, MIT, Cambridge
10. Andersen TG (1994) Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. Math Finance 4:75–102
11. Andersen TG (1996) Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. J Finance 51:169–204
12. Andersen TG, Benzoni L (2006) Do Bonds Span Volatility Risk in the US Treasury Market? A Specification Test for Affine Term Structure Models. Working Paper, KSM and Chicago FED, Chicago
13. Andersen TG, Benzoni L (2007) The Determinants of Volatility in the US Treasury market. Working Paper, KSM and Chicago FED, Chicago
14. Andersen TG, Benzoni L (2007) Realized volatility. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin (forthcoming)
15. Andersen TG, Benzoni L, Lund J (2002) An Empirical Investigation of Continuous-Time Equity Return Models. J Finance 57:1239–1284
16. Andersen TG, Benzoni L, Lund J (2004) Stochastic Volatility, Mean Drift and Jumps in the Short Term Interest Rate. Working Paper, Northwestern University, University of Minnesota, and Nykredit Bank, Copenhagen
17. Andersen TG, Bollerslev T (1998) Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. J Finance 53:219–265
18. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39:885–905
19. Andersen TG, Bollerslev T, Diebold FX (2004) Parametric and nonparametric volatility measurement. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
20. Andersen TG, Bollerslev T, Diebold FX (2007) Roughing It Up: Including Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility. Rev Econ Statist 89:701–720
21. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71:579–625
22. Andersen TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Ammer Econom Rev 93:38–62
23. Andersen TG, Bollerslev T, Dobrev D (2007) No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. J Econome 138:125–180
24. Andersen TG, Bollerslev T, Meddahi N (2004) Analytic Evaluation of Volatility Forecasts. Int Econ Rev 45:1079–1110
25. Andersen TG, Bollerslev T, Meddahi N (2005) Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. Econometrica 73:279–296
26. Andersen TG, Bondarenko O (2007) Construction and Interpretation of Model-Free Implied Volatility. Working Paper, KSM and UIC, Chicago
27. Andersen TG, Chung HJ, Sørensen BE (1999) Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study. J Econom 91:61–87
28. Andersen TG, Lund J (1997) Estimating continuous-time stochastic volatility models of the short term interest rate diffusion. J Econom 77:343–377
29. Ané T, Geman H (2000) Order Flow, Transaction Clock, and Normality of Asset Returns. J Finance 55:2259–2284
30. Ang A, Hodrick RJ, Xing Y, Zhang X (2006) The Cross-Section of Volatility and Expected Returns. J Finance 51:259–299

31. Ang A, Hodrick RJ, Xing Y, Zhang X (2008) High idiosyncratic volatility and low returns: international and further U.S. evidence. Finance Econ (forthcoming)
32. Bachelier L (1900) Théorie de la Spéculation. Annales de École Normale Supérieure 3, Gauthier-Villars, Paris. English translation: Cootner PH (ed) (1964) The Random Character of Stock Market Prices. MIT Press, Cambridge
33. Back K (1991) Asset Prices for General Processes. J Math Econ 20:371–395
34. Bakshi G, Cao C, Chen Z (1997) Empirical Performance of Alternative Option Pricing Models. J Finance 52:2003–2049
35. Bakshi G, Cao C, Chen Z (2002) Pricing and hedging long-term options. J Econom 94:277–318
36. Bakshi G, Kapadia N (2003) Delta-Hedged Gains and the Negative Market Volatility Risk Premium. Rev Finance Stud 16:527–566
37. Bakshi G, Kapadia N, Madan D (2003) Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. Rev Finance Stud 16:101–143
38. Balduzzi P, Elton EJ, Green TC (2001) Economic News and Bond Prices: Evidence from the US Treasury Market. J Finance Quant Anal 36:523–543
39. Bandi FM (2002) Short-term interest rate dynamics: a spatial approach. J Financ Econ 65:73–110
40. Bandi FM, Moise CE, Russel JR (2008) Market volatility, market frictions, and the cross section of stock returns. Working Paper, University of Chicago, and Case Western Reverse University, Cleveland
41. Bandi FM, Nguyen T (2003) On the functional estimation of jump-diffusion models. J Econom 116:293–328
42. Bandi FM, Phillips PCB (2002) Nonstationary Continuous-Time Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
43. Bandi FM, Phillips PCB (2003) Fully nonparametric estimation of scalar diffusion models. Econometrica 71:241–283
44. Bandi FM, Phillips PCB (2007) A simple approach to the parametric estimation of potentially nonstationary diffusions. J Econom 137:354–395
45. Bandi FM, Russell J (2006) Separating Microstructure Noise from Volatility. J Finance Econ 79:655–692
46. Bandi FM, Russell J (2007) Volatility. In: Birge J, Linetsky V (eds) Handbook of Financial Engineering. Elsevier, Amsterdam
47. Bansal R, Yaron A (2004) Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. J Finance 59:1481–1509
48. Bartlett MS (1938) The Characteristic Function of a Conditional Statistic. J Lond Math Soc 13:63–67
49. Barndorff-Nielsen OE, Hansen P, Lunde A, Shephard N (2007) Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. Working Paper, University of Aarhus, Aarhus. Stanford University, Nuffield College, Oxford
50. Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian Ornstein-Uhlenbeckbased models and some of their uses in financial economics. J R Stat Soc B 63:167–241
51. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. J R Stat Soc B 64:253–280
52. Barndorff-Nielsen OE, Shephard N (2002b) Estimating quadratic variation using realized variance. J Appl Econom 17:457–477
53. Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. J Finance Econom 2:1–37
54. Barndorff-Nielsen OE, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. J Finance Econom 4:1–30
55. Bates DS (1991) The Crash of ’87: Was It Expected? The Evidence from Options Markets. J Finance 46:1009–1044
56. Bates DS (1996) Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. Rev Finance Stud 9:69–107
57. Bates DS (2000) Post-’87 crash fears in the S&P 500 futures option market. J Econom 94:181–238
58. Bates DS (2006) Maximum Likelihood Estimation of Latent Affine Processes. Rev Finance Stud 19:909–965
59. Benzoni L (2002) Pricing Options under Stochastic Volatility: An Empirical Investigation. Working Paper, Chicago FED
60. Benzoni L, Collin-Dufresne P, Goldstein RS (2007) Explaining Pre- and Post-1987 Crash Prices of Equity and Options within a Unified General Equilibrium Framework. Working Paper, Chicago FED, UCB, and UMN, Minneapolis
61. Bibby BM, Jacobsen M, Sorensen M (2004) Estimating Functions for Discretely Sampled Diffusion-Type Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
62. Bikbov R, Chernov M (2005) Term Structure and Volatility: Lessons from the Eurodollar Markets. Working Paper, LBS, Deutsche Bank, New York
63. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. J Political Econ 81:637–654
64. Bladt M, Sørensen M (2007) Simple Simulation of Diffusion Bridges with Application to Likelihood Inference for Diffusions. Working Paper, University of Copenhagen, Copenhagen
65. Bollen NPB, Whaley RE (2004) Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? J Finance 59:711–753
66. Bollerslev T, Gibson M, Zhou H (2004) Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. Working Paper, Duke University, Federal Reserve Board, Washington D.C.
67. Bollerslev T, Jubinsky PD (1999) Equity Trading vol and Volatility: Latent Information Arrivals and Common Long-Run Dependencies. J Bus Econ Stat 17:9–21
68. Bollerslev T, Zhou H (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. J Econom 109:33–65
69. Bollerslev T, Zhou H (2007) Expected Stock Returns and Variance Risk Premia. Working Paper, Duke University and Federal Reserve Board, Washington D.C.
70. Bondarenko O (2004) Market price of variance risk and performance of hedge funds. Working Paper, UIC, Chicago
71. Brandt MW, Chapman DA (2003) Comparing Multifactor Models of the Term Structure. Working Paper, Duke University and Boston College, Chestnut Hill
72. Brandt MW, Diebold FX (2006) A No-Arbitrage Approach to Range-Based Estimation of Return Covariances and Correlations. J Bus 79:61–73
73. Brandt MW, Jones CS (2006) Volatility Forecasting with Range-Based EGARCH Models. J Bus Econ Stat 24:470–486
74. Brandt MW, Kavajecz KA (2004) Price Discovery in the US Trea-

sury Market: The Impact of Orderflow and Liquidity on the Yield Curve. J Finance 59:2623–2654

75. Brandt MW, Santa-Clara P (2002) Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. J Finance Econ 63:161–210
76. Breidt FJ, Crato N, de Lima P (1998) The detection and estimation of long memory in stochastic volatility. J Econom 83:325–348
77. Britten-Jones M, Neuberger A (2000) Option Prices, Implied Price Processes, and Stochastic Volatility. J Finance 55:839–866
78. Broadie M, Chernov M, Johannes MJ (2007) Model Specification and Risk Premia: Evidence from Futures Options. J Finance 62:1453–1490
79. Buraschi A, Jackwerth J (2001) The price of a smile: hedging and spanning in option markets. Rev Finance Stud 14:495–527
80. Campbell JY, Cochrane JH (1999) By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. J Polit Economy 107:205–251
81. Campbell JY, Viceira LM (2001) Who Should Buy Long-Term Bonds? Ammer Econ Rev 91:99–127
82. Carr P, Lee R (2007) Robust Replication of Volatility Derivatives. Working Paper, NYU and Uinversity of Chicago, Chicago
83. Carr P, Madan D (1998) Towards a theory of volatility trading. In: Jarrow R (ed) Volatility. Risk Publications, London
84. Carr P, Wu L (2004) Time-changed Lévy processes and option pricing. J Finance Econ 71:113–141
85. Carr P, Wu L (2007) Variance Risk Premia. Rev Finance Stud, forthcoming
86. Carrasco M, Chernov M, Florens JP, Ghysels E (2007) Efficient estimation of general dynamic models with a continuum of moment conditions. J Econom 140:529–573
87. Carrasco M, Florens JP (2000) Generalization of GMM to a continuum of moment conditions. Econom Theory 16:797–834
88. Chacko G, Viceira LM (2003) Spectral GMM estimation of continuous-time processes. J Econom 116:259–292
89. Chan KC, Karolyi GA, Longstaff FA, Sanders AB (1992) An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. J Finance 47:1209–1227
90. Chen J (2003) Intertemporal CAPM and the Cross-Section of Stock Return. Working Paper, USC, Los Angeles
91. Chen RR, Scott L (1993) Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. J Fixed Income 3:14–31
92. Cheridito P, Filipović D, Kimmel RL (2007) Market Price of Risk Specifications for Affine Models: Theory and Evidence. J Finance Econ, 83(1):123–170
93. Chernov M, Ghysels E (2002) A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. J Finance Econ 56:407–458
94. Chernov M, Gallant AR, Ghysels E, Tauchen G (1999) A New Class of Stochastic Volatility Models with Jumps: Theory and Estimation. Working Paper, London Business School, Duke University, University of Northern Carolina
95. Chernov M, Gallant AR, Ghysels E, Tauchen G (2003) Alternative models for stock price dynamics. J Econom 116:225–257
96. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
97. Chib S, Nardari F, Shephard N (2006) Analysis of high dimensional multivariate stochastic volatility models. J Econom 134:341–371
98. Christoffersen PF, Jacobs K (2004) The importance of the loss function in option valuation. J Finance Econ 72:291–318
99. Christoffersen PF, Jacobs K, Karoui L, Mimouni K (2007) Estimating Term Structure Models Using Swap Rates. Working Paper, McGill University, Montreal
100. Christoffersen PF, Jacobs K, Mimouni K (2006a) Models for S&P 500 Dynamics: Evidence from Realized Volatility, Daily Returns, and Option Prices. Working Paper, McGill University, Montreal
101. Christoffersen PF, Jacobs K, Wang Y (2006b) Option Valuation with Long-run and Short-run Volatility Components. Working Paper, McGill University, Montreal
102. Clark PK (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. Econometrica 41:135–156
103. Collin-Dufresne P, Goldstein RS (2002) Do Bonds Span the Fixed Income Markets? Theory and Evidence for Unspanned Stochastic Volatility. J Finance 57:1685–1730
104. Collin-Dufresne P, Goldstein R, Jones CS (2008) Identification of Maximal Affine Term Structure Models. J Finance 63(2):743–795
105. Collin-Dufresne P, Goldstein R, Jones CS (2007b) Can Interest Rate Volatility be Extracted from the Cross Section of Bond Yields? An Investigation of Unspanned Stochastic Volatility. Working Paper, UCB, USC, UMN, Minneapolis
106. Collin-Dufresne P, Solnik B (2001) On the Term Structure of Default Premia in the Swap and LIBOR Markets. J Finance 56:1095–1115
107. Comte F, Coutin L, Renault E (2003) Affine Fractional Stochastic Volatility Models with Application to Option Pricing. Working Paper, CIRANO, Montreal
108. Comte F, Renault E (1998) Long memory in continuous-time stochastic volatility models. Math Finance 8:291–323
109. Conley TG, Hansen LP, Luttmer EGJ, Scheinkman JA (1997) Short-term interest rates as subordinated diffusions. Rev Finance Stud 10:525–577
110. Corsi F (2003) A Simple Long Memory Model of realized Volatility. Working paper, University of Southern Switzerland, Lugano
111. Coval JD, Shumway T (2001) Expected Option Returns. J Finance 56:983–1009
112. Cox JC, Ingersoll JE, Ross SA (1985) An Intertemporal General Equilibrium Model of Asset Prices. Econometrica 53:363–384
113. Cox JC, Ingersoll JE, Ross SA (1985) A Theory of the Term Structure of Interest Rates. Econometrica 53:385–407
114. Dai Q, Singleton KJ (2000) Specification Analysis of Affine Term Structure Models. J Finance 55:1943–1978
115. Dai Q, Singleton KJ (2003) Term Structure Dynamics in Theory and Reality. Rev Finance Stud 16:631–678
116. Danielsson J (1994) Stochastic Volatility in Asset Prices: Estimation by Simulated Maximum Likelihood. J Econom 64:375–400
117. Danielsson J, Richard JF (1993) Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models. J Appl Econom 8:S153–S173
118. Deo R, Hurvich C (2001) On the Log Periodogram Regression Estimator of the Memory Parameter in Long Memory Stochastic Volatility Models. Econom Theory 17:686–710

119. Derman E, Kani I (1994) The volatility smile and its implied tree. Quantitative Strategies Research Notes, Goldman Sachs, New York
120. Diebold FX, Nerlove M (1989) The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. J Appl Econom 4:1–21
121. Diebold FX, Strasser G (2007) On the Correlation Structure of Microstructure Noise in Theory and Practice. Working Paper, University of Pennsylvania, Philadelphia
122. Dobrev D (2007) Capturing Volatility from Large Price Moves: Generalized Range Theory and Applications. Working Paper, Federal Reserve Board, Washington D.C.
123. Driessen J, Maenhout P, Vilkov G (2006) Option-Implied Correlations and the Price of Correlation Risk. Working Paper, University of Amsterdam and INSEAD, Amsterdam
124. Duarte J (2004) Evaluating An Alternative Risk Preference in Affine Term Structure Models. Rev Finance Stud 17:370–404
125. Duarte J (2007) The Causal Effect of Mortgage Refinancing on Interest-Rate Volatility: Empirical Evidence and Theoretical Implications. Rev Finance Stud, forthcoming
126. Duarte J, Jones CS (2007) The Price of Market Volatility Risk. Working Paper, University of Washington and USC, Washington
127. Duffee GR (2002) Term Premia and Interest Rate Forecasts in Affine Models. J Finance 57:405–443
128. Duffee G, Stanton R (2008) Evidence on simulation inference for near unit-root processes with implications for term structure estimation. J Finance Econom 6:108–142
129. Duffie D, Kan R (1996) A yield-factor model of interest rates. Math Finance 6:379–406
130. Duffie D, Pan J, Singleton KJ (2000) Transform Analysis and Asset Pricing for Affine Jump-Diffusions. Econometrica 68:1343–1376
131. Duffie D, Singleton KJ (1993) Simulated Moments Estimation of Markov Models of Asset Prices. Econometrica 61:929–952
132. Duffie D, Singleton KJ (1997) An Econometric Model of the Term Structure of Interest-Rate Swap Yields. J Finance 52:1287–1321
133. Dumas B, Fleming J, Whaley RE (1996) Implied Volatility Functions: Empirical Tests. J Finance 53:2059–2106
134. Dupire B (1994) Pricing with a smile. Risk 7:18–20
135. Elerian O, Chib S, Shephard N (2001) Likelihood Inference for Discretely Observed Nonlinear Diffusions. Econometrica 69:959–994
136. Engle RF (2002) New frontiers for ARCH models. J Appl Econom 17:425–446
137. Engle RF, Gallo GM (2006) A multiple indicators model for volatility using intra-daily data. J Econom 131:3–27
138. Engle RF, Ghysels E, Sohn B (2006) On the Economic Sources of Stock Market Volatility. Working Paper, NYU and UNC, Chapel Hill
139. Engle RF, Rangel JG (2006) The Spline-GARCH Model for Low Frequency Volatility and Its Global Macroeconomic Causes. Working Paper, NYU, New York
140. Eraker B (2001) MCMC Analysis of Diffusions with Applications to Finance. J Bus Econ Stat 19:177–191
141. Eraker B (2004) Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. J Finance 59:1367–1404
142. Eraker B, Johannes MS, Polson N (2003) The Impact of Jumps in Volatility and Returns. J Finance 58:1269–1300
143. Fan R, Gupta A, Ritchken P (2003) Hedging in the Possible Presence of Unspanned Stochastic Volatility: Evidence from Swaption Markets. J Finance 58:2219–2248
144. Fiorentina G, Sentana E, Shephard N (2004) Likelihood-Based Estimation of Latent Generalized ARCH Structures. Econometrica 72:1481–1517
145. Fisher M, Gilles C (1996) Estimating exponential-affine models of the term structure. Working Paper, Atlanta FED, Atlanta
146. Fleming MJ, Remolona EM (1999) Price Formation and Liquidity in the US Treasury Market: The Response to Public Information. J Finance 54:1901–1915
147. Forsberg L, Ghysels E (2007) Why Do Absolute Returns Predict Volatility So Well? J Finance Econom 5:31–67
148. French KR, Schwert GW, Stambaugh RF (1987) Expected stock returns and volatility. J Finance Econ 19:3–29
149. Fridman M, Harris L (1998) A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models. J Bus Econ Stat 16:284–291
150. Gallant AR, Hsieh DA, Tauchen GE (1997) Estimation of Stochastic Volatility Models with Diagnostics. J Econom 81:159–192
151. Gallant AR, Hsu C, Tauchen GE (1999) Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance. Rev Econ Stat 81:617–631
152. Gallant AR, Long JR (1997) Estimating stochastic differential equations efficiently by minimum chi-squared. Biometrika 84:125–141
153. Gallant AR, Rossi PE, Tauchen GE (1992) Stock Prices and Volume. Rev Finance Stud 5:199–242
154. Gallant AR, Tauchen GE (1996) Which Moments to Match. Econ Theory 12:657–681
155. Gallant AR, Tauchen G (1998) Reprojecting Partially Observed Systems With Application to Interest Rate Diffusions. J Ammer Stat Assoc 93:10–24
156. Garcia R, Lewis MA, Pastorello S, Renault E (2001) Estimation of Objective and Risk-neutral Distributions based on Moments of Integrated Volatility, Working Paper. Université de Montréal, Banque Nationale du Canada, Università di Bologna, UNC
157. Garman MB, Klass MJ (1980) On the Estimation of Price Volatility From Historical Data. J Bus 53:67–78
158. Ghysels E, Harvey AC, Renault E (1996) Stochastic Volatility. In: Maddala GS, Rao CR (eds) Handbook of Statistics, vol 14. North Holland, Amsterdam
159. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get the Most Out of Returns Data Sampled at Different Frequencies. J Econom 131:59–95
160. Glosten LR, Jagannathan R, Runkle D (1993) On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801
161. Gong FF, Remolona EM (1996) A three-factor econometric model of the US term structure. Working paper, Federal Reserve Bank of New York, New York
162. Gouriéroux C, Monfort A, Renault E (1993) Indirect Inference. J Appl Econom 8:S85–S118
163. Guo H, Neely CJ, Higbee J (2007) Foreign Exchange Volatility Is Priced in Equities. Finance Manag, forthcoming
164. Han B (2007) Stochastic Volatilities and Correlations of Bond Yields. J Finance 62:1491–1524
165. Hansen PR, Lunde A (2006) Realized Variance and Market Microstructure Noise. J Bus Econ Stat 24:127–161

166. Harvey AC (1998) Long memory in stochastic volatility. In: Knight J, Satchell S (eds) Forecasting Volatility in Financial Markets. Butterworth-Heinemann, London
167. Harvey AC, Ruiz E, Shephard N (1994) Multivariate Stochastic Variance Models. Rev Econ Stud 61:247–264
168. Harvey AC, Shephard N (1996) Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. J Bus Econ Stat 14:429–434
169. Heidari M, Wu L (2003) Are Interest Rate Derivatives Spanned by the Term Structure of Interest Rates? J Fixed Income 13:75–86
170. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Finance Stud 6:327–343
171. Ho M, Perraudin W, Sørensen BE (1996) A Continuous Time Arbitrage Pricing Model with Stochastic Volatility and Jumps. J Bus Econ Stat 14:31–43
172. Huang X, Tauchen G (2005) The relative contribution of jumps to total price variation. J Finance Econom 3:456–499
173. Huang J, Wu L (2004) Specification Analysis of Option Pricing Models Based on Time-Changed Levy Processes. J Finance 59:1405–1440
174. Hull J, White A (1987) The Pricing of Options on Assets with Stochastic Volatilities. J Finance 42:281–300
175. Hurvich CM, Soulier P (2007) Stochastic Volatility Models with Long Memory. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer,Berlin
176. Jung J (2006) Vexed by variance. Risk August
177. Jackwerth JC, Rubinstein M (1996) Recovering Probability Distributions from Option Prices. J Finance 51:1611–1631
178. Jacobs K, Karoui L (2007) Conditional Volatility in Affine Term Structure Models: Evidence from Treasury and Swap Markets. Working Paper, McGill University, Montreal
179. Jacobs K, Li X (2008) Modeling the Dynamics of Credit Spreads with Stochastic Volatility. Manag Sci 54:1176–1188
180. Jacquier E, Polson NG, Rossi PE (1994) Bayesian Analysis of Stochastic Volatility Models. J Bus Econ Stat 12:371–389
181. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J Econom 122:185–212
182. Jagannathan R, Kaplin A, Sun S (2003) An evaluation of multi-factor CIR models using LIBOR, swap rates, and cap and swaption prices. J Econom 116:113–146
183. Jarrow R, Li H, Zhao F (2007) Interest Rate Caps "Smile" Too! But Can the LIBOR Market Models Capture the Smile? J Finance 62:345–382
184. Jiang GJ, Knight JL (2002) Efficient Estimation of the Continuous Time Stochastic Volatility Model via the Empirical Characteristic Function. J Bus Econ Stat 20:198–212
185. Jiang GJ, Tian YS (2005) The Model-Free Implied Volatility and Its Information Content. Rev Finance Stud 18:1305–1342
186. Johannes MS, Polson N (2003) MCMC Methods for Continuous-Time Financial Econometrics. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
187. Johannes MS, Polson N (2006) Particle Filtering. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin
188. Johnson H, Shanno D (1987) Option Pricing when the Variance Is Changing. J Finance Quant Anal 22:143–152
189. Jones CS (2003) Nonlinear Mean Reversion in the Short-Term Interest Rate. Rev Finance Stud 16:793–843
190. Jones CS (2003) The dynamics of stochastic volatility: evidence from underlying and options markets. J Econom 116:181–224
191. Joslin S (2006) Can Unspanned Stochastic Volatility Models Explain the Cross Section of Bond Volatilities? Working Paper, MIT, Cambridge
192. Joslin S (2007) Pricing and Hedging Volatility Risk in Fixed Income Markets. Working Paper, MIT, Cambridge
193. Kim SN, Shephard N, Chib S (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. Rev Econ Stud 65:361–393
194. Lamoureux CG, Lastrapes WD (1994) Endogenous Trading vol and Momentum in Stock-Return Volatility. J Bus Econ Stat 14:253–260
195. Lee SS, Mykland PA (2006) Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. Working Paper, Georgia Institute of Technology and University of Chicago, Chicago
196. Li H, Wells MT, Yu CL (2006) A Bayesian Analysis of Return Dynamics with Lévy Jumps. Rev Finance Stud, forthcoming
197. Li H, Zhao F (2006) Unspanned Stochastic Volatility: Evidence from Hedging Interest Rate Derivatives. J Finance 61:341–378
198. Liesenfeld R (1998) Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading Volume. J Bus Econ Stat 16:101–109
199. Liesenfeld R (2001) A Generalized Bivariate Mixture Model for Stock Price Volatility and Trading Volume. J Econom 104:141–178
200. Liesenfeld R, Richard J-F (2003) Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. J Empir Finance 10:505–531
201. Litterman R, Scheinkman JA (1991) Common Factors Affecting Bond Returns. J Fixed Income 1:54–61
202. Liu C, Maheu JM (2007) Forecasting Realized Volatility: A Bayesian Model Averaging Approach. Working Paper, University of Toronto, Toronto
203. Lo AW (1988) Maximum likelihood estimation of generalized Itô processes with discretely-sampled data. Econom Theory 4:231–247
204. Longstaff FA, Schwartz ES (1992) Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. J Finance 47:1259–1282
205. McAleer M, Medeiros MC (2007) Realized Volatility: A Review. Econom Rev, forthcoming
206. McFadden D (1989) A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. Econometrica 57:995–1026
207. McQueen G, Vorkink K (2004) Whence GARCH? A Preference-Based Explanation for Conditional Volatility. Rev Finance Stud 17:915–949
208. Meddahi N (2001) An Eigenfunction Approach for Volatility Modeling. Working Paper, Imperial College, London
209. Meddahi N (2002) A Theoretical Comparison Between Integrated and Realized Volatility. J Appl Econom 17:475–508
210. Meddahi N (2002) Moments of Continuous Time Stochastic Volatility Models. Working Paper, Imperial College, London
211. Melino A, Turnbull SM (1990) Pricing foreign currency options with stochastic volatility. J Econom 45:239–265

212. Merton RC (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. Rev Econ Stat 51:247–257
213. Merton RC (1973) An Intertemporal Capital Asset Pricing Model. Econometrica 41:867–887
214. Merton RC (1976) Option pricing when underlying stock returns are discontinuous. J Finance Econs 3:125–144
215. Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. J Finance Econ 8:323–361
216. Mizrach B (2006) The Enron Bankruptcy: When Did The Options Market Lose Its Smirk. Rev Quant Finance Acc 27:365–382
217. Mizrach B (2007) Recovering Probabilistic Information From Options Prices and the Underlying. In: Lee C, Lee AC (eds) Handbook of Quantitative Finance. Springer, New York
218. Nelson DB (1991) Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica 59:347–370
219. Nelson DB, Foster DP (1994) Asymptotic Filtering Theory for Univariate ARCH Models. Econometrica 62:1–41
220. Pan J (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. J Finance Econ 63:3–50
221. Parkinson M (1980) The Extreme ValueMethod for Estimating the Variance of the Rate of Return. J Bus 53:61–65
222. Pasquariello P, Vega C (2007) Informed and Strategic Order Flow in the Bond Markets. Rev Finance Stud 20:1975–2019
223. Pearson ND, Sun TS (1994) Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model. J Finance 49:1279–1304
224. Pedersen AR (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand J Stat 22:55–71
225. Pennacchi GG (1991) Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data. Rev Finance Stud 4:53–86
226. Piazzesi M (2003) Affine Term Structure Models. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
227. Piazzesi M (2005) Bond Yields and the Federal Reserve. J Political Econ 113:311–344
228. Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filter. J Ammer Stat Assoc 94:590–599
229. Protter P (1992) Stochastic Integration and Differential Equations: A New Approach. Springer, New York
230. Renault E (1997) Econometric models of option pricing errors. In: Kreps D, Wallis K (eds) Advances in Economics and Econometrics, Seventh World Congress. Cambridge University Press, New York, pp 223–278
231. Richardson M, Smith T (1994) A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information. J Financ Quant Anal 29:101–116
232. Rosenberg B (1972) The Behavior of Random Variables with Nonstationary Variance and the Distribution of Security Prices. working Paper, UCB, Berkley
233. Rubinstein M (1994) Implied Binomial Trees. J Financ 49:771–818
234. Santa-Clara P (1995) Simulated likelihood estimation of diffusions with an application to the short term interest rate. Dissertation, INSEAD
235. Schaumburg E (2005) Estimation of Markov processes with Levy type generators. Working Paper, KSM, Evanston
236. Schwert GW (1989) Why Does Stock Market Volatility Change Over Time? J Financ 44:1115–1153
237. Schwert GW (1990) Stock Volatility and the Crash of ’87. Rev Financ Stud 3:77–102
238. Scott LO (1987) Option Pricing when the Variance Changes Randomly: Theory, Estimation and an Application. J Financ Quant Anal 22:419–438
239. Shephard N (1996) Statistical Aspects of ARCH and Stochastic Volatility Models. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London, pp 1–67
240. Shephard N (2004) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
241. Shiller RJ (1981) Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? Ammer Econ Rev 71:421–436
242. Singleton KJ (2001) Estimation of affine asset pricing models using the empirical characteristic function. J Econom 102:111–141
243. Smith AA Jr (1993) Estimating Nonlinear Time-series Models using Simulated Vector Autoregressions. J Appl Econom 8:S63–S84
244. Stein JC (1989) Overreactions in the Options Market. J Finance 44:1011–1023
245. Stein EM, Stein JC (1991) Stock price distributions with stochastic volatility: an analytic approach. Rev Finance Stud 4:727–752
246. Tauchen GE (2005) Stochastic Volatility in General Equilibrium. Working Paper, Duke, University Durham
247. Tauchen GE, Pitts M (1983) The Price Variability-Volume Relationship on Speculative Markets. Econometrica 51:485–505
248. Tauchen GE, Zhou H (2007) Realized Jumps on Financial Markets and Predicting Credit Spreads. Working Paper, Duke University and Board of Governors, Washington D.C.
249. Taylor SJ (1986) Modeling Financial Time Series. Wiley, Chichester
250. Thompson S (2004) Identifying Term Structure Volatility from the LIBOR-Swap Curve. Working Paper, Harvard University, Boston
251. Todorov V (2006) Variance Risk Premium Dynamics. Working Paper, KSM, Evanston
252. Todorov V (2006) Estimation of Continuous-time Stochastic Volatility Models with Jumps using High-Frequency Data. Working Paper, KSM, Evanston
253. Trolle AB, Schwartz ES (2007) Unspanned stochastic volatility and the pricing of commodity derivatives. Working Paper, Copenhagen Business School and UCLA, Copenhagen
254. Trolle AB, Schwartz ES (2007) A general stochastic volatility model for the pricing of interest rate derivatives. Rev Finance Stud, forthcoming
255. Vasicek OA (1977) An equilibrium characterization of the term structure. J Finance Econ 5:177–188
256. Wiggins JB (1987) Option Values under Stochastic Volatility: Theory and Empirical Estimates. J Finance Econ 19:351–372
257. Whaley RE (1993) Derivatives on Market Volatility: Hedging Tools Long Overdue. J Deriv 1:71–84
258. Wright J, Zhou H (2007) Bond Risk Premia and Realized Jump Volatility. Working Paper, Board of Governors, Washington D.C.

259. Yang D, Zhang Q (2000) Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. J Bus 73:477–491
260. Zhang BY, Zhou H, Zhu H (2005) Explaining Credit Default Swap Spreads with the Equity Volatility and Jump Risks of Individual Firms. Working Paper, Fitch Ratings, Federal Reserve Board, BIS
261. Zhang L (2007) What you don't know cannot hurt you: On the detection of small jumps. Working Paper, UIC, Chicago
262. Zhang L, Mykland PA, Aït-Sahalia Y (2005) A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data. J Ammer Stat Assoc 100:1394–1411

Books and Reviews

Asai M, McAleer M, Yu J (2006) Multivariate Stochastic Volatility: A Review. Econom Rev 25:145–175

Bates DS (2003) Empirical option pricing: a retrospection. J Econom 116:387–404

Campbell JY, Lo AW, MacKinlay AC (1996) The Econometrics of Financial Markets. Princeton University Press, Princeton

Chib S, Omori Y, Asai M (2007) Multivariate Stochastic Volatility. Working Paper, Washington University

Duffie D (2001) Dynamic Asset Pricing Theory. Princeton University Press, Princeton

Gallant AR, Tauchen G (2002) Simulated Score Methods and Indirect Inference for Continuous-time Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Garcia R, Ghysels E, Renault E (2003) The Econometrics of Option Pricing. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Gouriéroux C, Jasiak J (2001) Financial Econometrics. Princeton University Press, Princeton

Johannes MS, Polson N (2006) Markov Chain Monte Carlo. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin

Jungbacker B, Koopman SJ (2007) Parameter Estimation and Practical Aspect of Modeling Stochastic Volatility. Working Paper, Vrije Universiteit, Amsterdam

Mykland PA (2003) Option Pricing Bounds and Statistical Uncertainty. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming

Renault E (2007) Moment-Based Estimation of Stochastic Volatility Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin

Singleton KJ (2006) Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment. Princeton University Press, Princeton

System Dynamics and Its Contribution to Economics and Economic Modeling

MICHAEL J. RADZICKI
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary

Stock Stocks, which are sometimes referred to as "levels" or "states", accumulate (i. e., sum up) the information or material that flows into and out of them. Stocks are thus responsible for decoupling flows, creating delays, preserving system memory, and altering the time shape of flows.

Flow Flows of information or material enter and exit a system's stocks and, in so doing, create a system's dynamics. Stated differently, the net flow into or out of a stock is the stock's rate of change. When human decision making is represented in a system dynamics model, it appears in the system's flow equations. Mathematically, a system's flow equations are ordinary differential equations and their format determines whether or not a system is linear or nonlinear.

Feedback Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. When the return of this information reinforces a system's behavior, the loop is said to be positive. Positive loops are responsible for the exponential growth of a system over time. Negative feedback loops represent goal seeking behavior in complex systems. When a negative loop detects a gap between the amount of information or material in a system's stock and the desired amount of information or material, it initiates corrective action. If this corrective action is not significantly delayed, the system will smoothly adjust to its goal. If the corrective action is delayed, however, the system can overshoot or undershoot its goal and the system can oscillate.

Full information maximum likelihood with optimal filtering FIMLOF is a sophisticated technique for estimating the parameters of a system dynamics model, while simultaneously fitting its output to numerical data. Its intellectual origins can be traced to control engineering and the work of Fred Schwepe. David Peterson pioneered a method for adapting FIMLOF for use in system dynamics modeling.

Definition of the Subject

System dynamics is a computer modeling method that has its intellectual origins in control engineering, management science, and digital computing. It was originally created as a tool to help managers better understand and control corporate systems. Today it is applied to problems in a wide variety of academic disciplines, including economics. Of note is that system dynamics models often generate behavior that is both counterintuitive and at odds with traditional economic theory. Historically, this has caused many system dynamics models to be evaluated critically, especially by some economists. However, today economists from several schools of economic thought are beginning to use system dynamics, as they have found it useful for incorporating their nontraditional ideas into formal models.

Introduction

System dynamics is a computer simulation modeling methodology that is used to analyze complex nonlinear dynamic feedback systems for the purposes of generating insight and designing policies that will improve system performance. It was originally created in 1957 by Jay W. Forrester of the Massachusetts Institute of Technology as a method for building computer simulation models of problematic behavior within corporations. The models were used to design and test policies aimed at altering a corporation's structure so that its behavior would improve and become more robust. Today, system dynamics is applied to a large variety of problems in a multitude of academic disciplines, including economics.

System dynamics models are created by identifying and linking the relevant pieces of a system's structure and simulating the behavior generated by that structure. Through an iterative process of structure identification, mapping, and simulation a model emerges that can explain (mimic) a system's problematic behavior and serve as a vehicle for policy design and testing.

From a system dynamics perspective a system's structure consists of stocks, flows, feedback loops, and limiting factors. Stocks can be thought of as bathtubs that accumulate/de-cumulate a system's flows over time. Flows can be thought of as pipe and faucet assemblies that fill or drain the stocks. Mathematically, the process of flows accumulating/de-cumulating in stocks is called integration. The integration process creates all dynamic behavior in the world be it in a physical system, a biological system, or a socioeconomic system. Examples of stocks and flows in economic systems include a stock of inventory and its inflow of production and its outflow of sales, a stock of the book value of a firm's capital and its inflow of investment spending and its outflow of depreciation, and a stock of employed labor and its inflow of hiring and its outflow of labor separations.

Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. Information travels from a stock back to its flow(s) either directly or indirectly, and this movement of information causes the system's faucets to open more, close a bit, close all the way, or stay in the same place. Every feedback loop has to contain at least one stock so that a simultaneous equation situation can be avoided and a model's behavior can be revealed recursively. Loops with a single stock are termed minor, while loops containing more than one stock are termed major.

Two types of feedback loops exist in system dynamics modeling: positive loops and negative loops. Generally speaking, positive loops generate self-reinforcing behavior and are responsible for the growth or decline of a system. Any relationship that can be termed a virtuous or vicious circle is thus a positive feedback loop. Examples of positive loops in economic systems include path dependent processes, increasing returns, speculative bubbles, learning-by-doing, and many of the relationships found in macroeconomic growth theory. Forrester [12], Radzicki and Sterman [46], Moxnes [32], Sterman (Chap. 10 in [55]), Radzicki [44], Ryzhenkov [49], and Weber [58] describe system dynamics models of economic systems that possess dominant positive feedback processes.

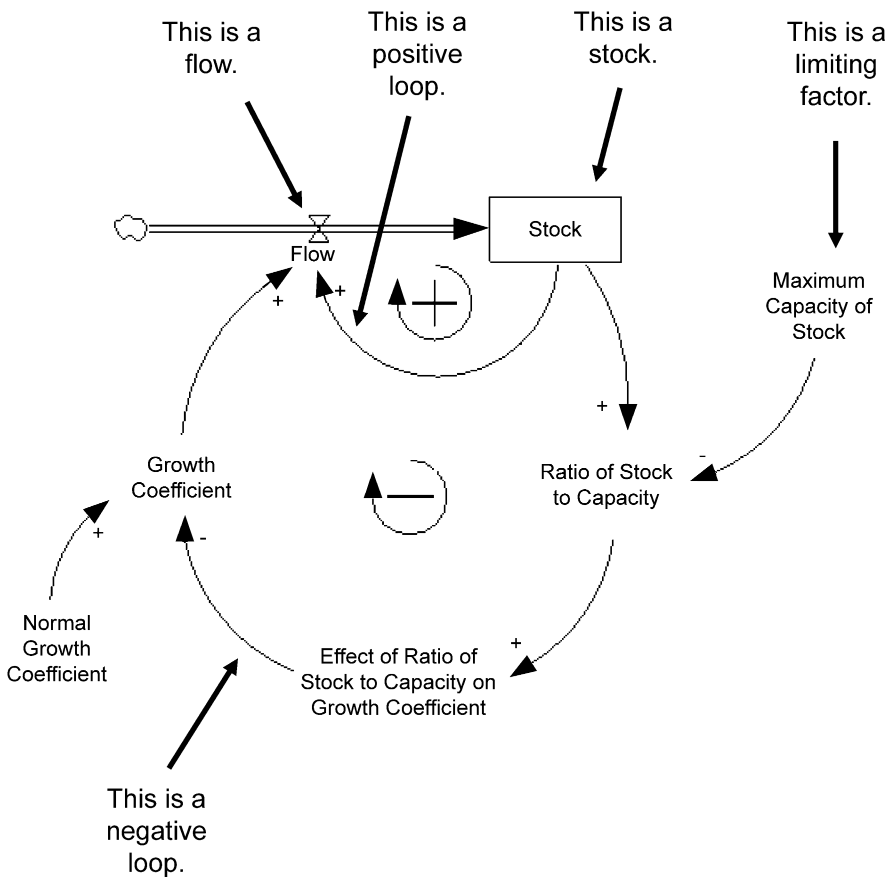
Negative feedback loops generate goal-seeking behavior and are responsible for both stabilizing systems and causing them to oscillate. When a negative loop detects a gap between a stock and its goal it initiates corrective action aimed at closing the gap. When this is accomplished without a significant time delay, a system will adjust smoothly to its goal. On the other hand, if there are significant time lags in the corrective actions of a negative loop, it can overshoot or undershoot its goal and cause the system to oscillate. Examples of negative feedback processes in economic systems include equilibrating mechanisms ("auto-pilots") such as simple supply and demand relationships, stock adjustment models for inventory control, any purposeful behavior, and many of the relationships found in macroeconomic business cycle theory. Meadows [27], Mass [26], Low [23], Forrester [12], and Sterman [54] provide examples of system dynamics models that generate cyclical behavior at the macro-economic and micro-economic levels.

From a system dynamics point of view, positive and negative feedback loops fight for control of a system's behavior. The loops that are dominant at any given time determine a system's time path and, if the system is nonlinear, the dominance of the loops can change over time as the system's stocks fill and drain. From this perspective, the dynamic behavior of any economy – that is, the interactions between the trend and the cycle in an economy over time – can be explained as a fight for dominance between the economy's most significant positive and negative feedback loops.

In system dynamics modeling, stocks are usually conceptualized as having limits. That is, stocks are usually seen as being unable to exceed or fall below certain maximum and minimum levels. Indeed, an economic model that can generate, say, either an infinite and/or a negative workforce would be seen as severely flawed by a system dynamicist. As such, when building a model system dynamicists search for factors that may limit the amount of material or information that the model's stocks can accumulate. Actual socioeconomic systems possess many limiting factors including physical limits (e. g., the number of widgets a machine can produce per unit of time), cognitive limits (e. g., the amount of information an economic agent can remember and act upon), and financial limits (e. g., the maximum balance allowed on a credit card). When limiting factors are included in a system dynamics model, the system's approach to these factors must be described. Generally speaking, this is accomplished with nonlinear relationships. Figure 1 presents a simple system dynamics model that contains examples of all of the components of system structure described above.

Types of Dynamic Simulation

From a system dynamics point of view, solving a dynamic model – any dynamic model – means determining how much material or information has accumulated in each of a system's stocks at every point in time. This can be accomplished in one of two ways – analytically or via sim-


This is a flow.
This is a positive loop.
This is a stock.
This is a limiting factor.
Stock
Flow
+
+
Maximum Capacity of Stock
+
-
Growth Coefficient
Ratio of Stock to Capacity
+
-
Normal Growth Coefficient
+
Effect of Ratio of Stock to Capacity on Growth Coefficient
This is a negative loop.

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 1
Simple system dynamics model containing examples of all components of system structure

ulation. Linear dynamic models can be solved either way. Nonlinear models, except for a few special cases, can only be solved via simulation.

Simulated solutions to dynamic systems can be attained from either a continuous (analog) computer or a discrete (digital) computer. Understanding the basic ideas behind the two approaches is necessary for understanding how economic modeling is undertaken with system dynamics.

In the real world, of course, time unfolds continuously. Yet, devising a way to mimic this process on a machine is a bit tricky. On an analog computer, the continuous flow of economic variables in and out of stocks over time is mimicked by the continuous flow of some physical substance such as electricity or water. A wonderful example of the later case is the Phillips Machine, which simulates an orthodox Keynesian economy (essentially the IS-LM model) with flows of colored water moving through pipes and accumulating in tanks. Barr [2] provides a vivid description of the history and restoration of the Phillips Machine.

On a digital computer, the continuous flow of economic variables in and out of stocks over time is approximated by specifying the initial amount of material or information in a system's stocks, breaking simulated time into small increments, inching simulated time forward by one of these small increments, calculating the amount of material or information that flowed into and out of the system's stocks during this small interval, and then repeating. The solution to the system will always be approximate because the increment of time cannot be made infinitesimally small and thus simulated time cannot be made perfectly continuous. In fact, on a digital computer a trade-off exists between round-off error and integration error. If the increment of time is made too large, the approximate solution can be poor due to integration error. If the increment of time is made too small, the approximate solution can be ruined due to round-off error.

In system dynamics modeling the "true" behavior of the underlying system is conceptualized to unfold over continuous time. As such, mathematically, a system dynamics model is an ordinary differential equation model. To approximate the solution to a continuous time ordinary differential equation model on a digital (discrete) computer, however, difference equations are used. Unlike traditional difference equation modeling in economics, in which the increment of time is chosen to match economic data (typically a quarter or a year), the increment of time in system dynamics modeling is chosen to yield a solution that is accurate enough for the problem at hand, yet avoids the problems associated with significant round-off and integration error.

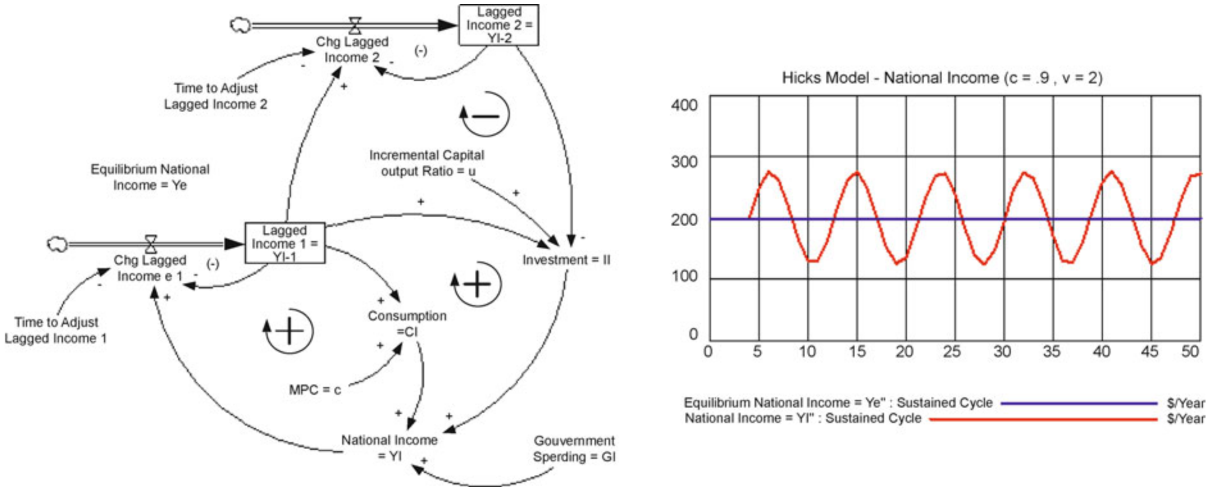
The use of difference equations to approximate the underlying differential equations represented by a system dynamics model provides another interesting option when it comes to economic modeling. Since many well known dynamic economic models have been created with difference equations, they can be recast in a system dynamics format by using the difference equations in the system dynamics software literally as difference equations, and not as a tool to approximate the underlying continuous time system. Although doing this deviates from the original ideas embodied in the system dynamics paradigm, it is occasionally done when a modeler feels that analyzing a difference equation model in a system dynamics format will yield some additional insight.

Translating Existing Economic Models into a System Dynamics Format

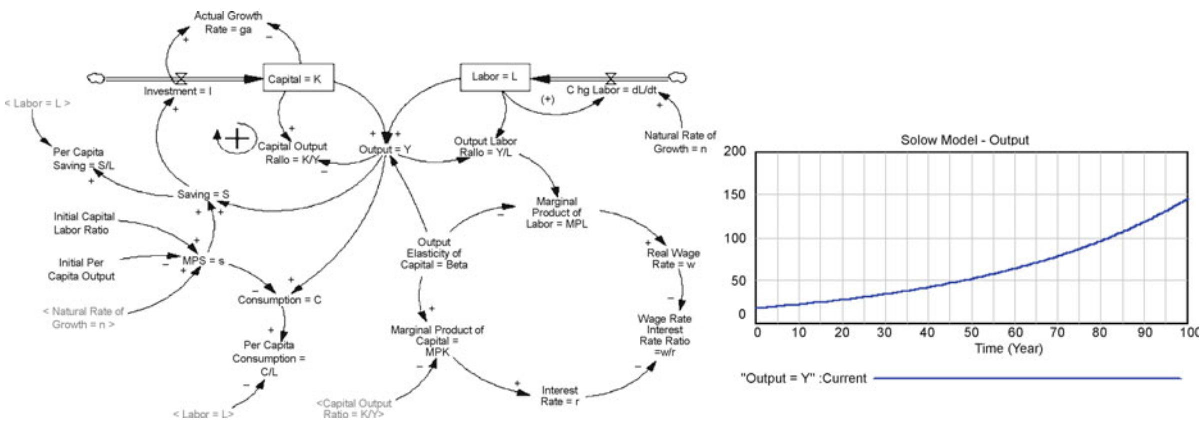
There are three principle ways that system dynamics is used for economic modeling. The first involves translating an existing economic model into a system dynamics format, while the second involves creating an economic model from scratch by following the rules and guidelines of the system dynamics paradigm. Forrester [7], Richardson and Pugh [47], Radzicki [42], and Sterman [55] provide extensive details about these rules and guidelines. The former approach is valuable because it enables well-known economic models to be represented in a common format, which makes comparing and contrasting their assumptions, concepts, structures, behaviors, etc., fairly easy. The latter approach is valuable because it usually yields models that are more realistic and that produce results that are "counterintuitive" [11] and thus thought-provoking.

The third way that system dynamics can be used for economic modeling is a "hybrid" approach in which a well known economic model is translated into a system dynamics format, critiqued, and then improved by modifying it so that it more closely adheres to the principles of system dynamics modeling. This approach attempts to blend the advantages of the first two approaches, although it is more closely related to the former.

Generally speaking, existing economic models that can be translated into a system dynamics format can be divided into four categories: written, static (mathematical), difference equation, and ordinary differential equation. Existing economic models that have been created in either a difference equation or an ordinary differential equation format can be translated into system dynamics in a fairly straight-forward manner. For example, Fig. 2 presents Sir John Hicks' [21] Multiplier-Accelerator difference equation model in a system dynamics format and Fig. 3 presents the Robert Solow's [52] ordinary differential equation growth model in a system dynamics format.


Lagged Income 2 = YI-2
Chg Lagged Income 2
(-)
Time to Adjust Lagged Income 2
Equilibrium National Income = Ye
Incremental Capital output Ratio = u
Lagged Income 1 = YI-1
Chg Lagged Income e 1
(-)
Investment = II
Time to Adjust Lagged Income 1
Consumption =CI
MPC = c
National Income = YI
Gouvemment Sperding = GI
Hicks Model - National Income (c = .9 , v = 2)
400
300
200
100
0
0
5
10
15
20
25
30
35
40
45
50
Equilibrium National Income = Ye" : Sustained Cycle
$/Year
National Income = YI" : Sustained Cycle
$/Year

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 2
System dynamics representation of John Hicks' multiplier-accelerator difference equation model


Actual Growth Rate = ga
Capital = K
Investment = I
< Labor = L >
Per Capita Saving = S/L
Capital Output Rallo = K/Y
Output = Y
Labor = L
C hg Labor = dL/dt
(+)
Output Labor Rallo = Y/L
Natural Rate of Growth = n
Saving = S
Initial Capital Labor Ratio
Initial Per Capita Output
MPS = s
< Natural Rate of Growth = n >
Consumption = C
Per Capita Consumption = C/L
< Labor = L>
Marginal Product of Labor = MPL
Output Elasticity of Capital = Beta
Real Wage Rate = w
Wage Rate Interest Rate Ratio =w/r
Marginal Product of Capital = MPK
<Capital Output Ratio = K/Y>
Interest Rate = r
Solow Model - Output
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
Time (Year)
"Output = Y" :Current

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 3
System dynamics representation of Robert Solow's ordinary differential equation growth model

Translating existing static and written economic models and theories into a system dynamics format is a more formidable task. Written models and theories are often dynamic, yet are described without mathematics. Static models and theories are often presented with mathematics, but lack equations that describe the dynamics of any adjustment processes they may undergo. As such, system dynamicists must devise equations that capture the dynamics being described by the written word or that reveal the adjustment processes that take place when a static system moves from one equilibrium point to another.

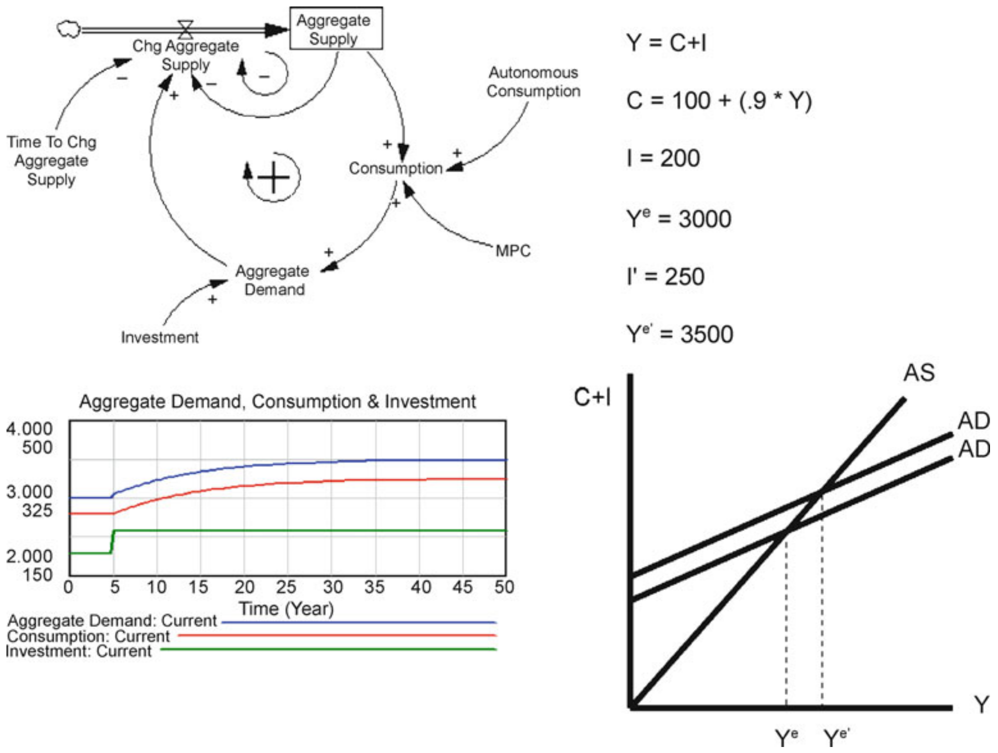
An interesting example of a system dynamics model that was created from a written economic model is Barry Richmond's [48] model of Adam Smith's *Wealth of Nations*. This model was created principally from Robert Heilbronner's [20] written description of Smith's economic system. A classic example of a static model that has been translated into a system dynamics format is a simple two sector Keynesian cross model, as is shown in Fig. 4.

Improving Existing Economic Models with System Dynamics

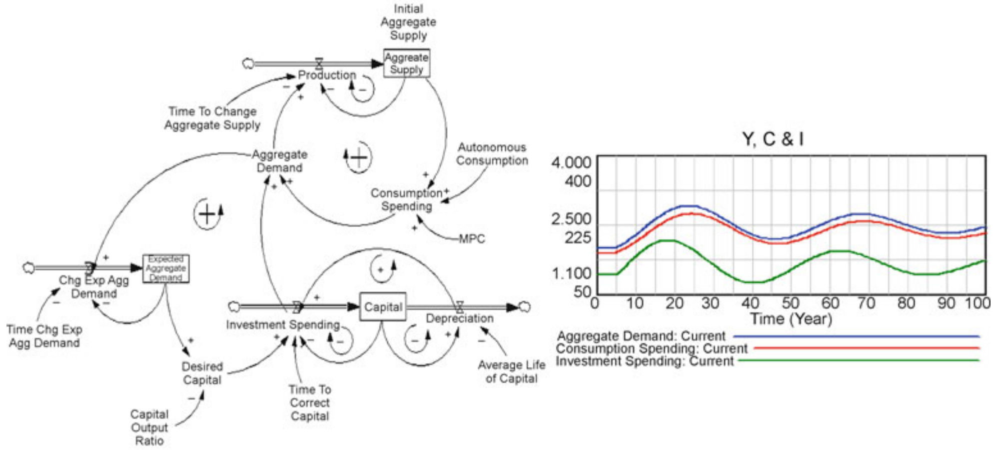
The simple two sector Keynesian cross model presented in Fig. 4 is an example of a well known economic model that can be improved after it has been translated into a system dynamics format. More specifically, in this example the flow of investment spending in the model does not accumulate anywhere. This violates good system dynamics modeling practice and can be fixed. Figure 5 presents the improved version of the Keynesian Cross model, which now more closely adheres to the system dynamics paradigm. Other well known examples of classic economics models that have been improved after they have been translated into a system dynamics format and made to conform more closely with good system dynamics modeling practice include the cobweb model [27], Sir John Hicks' multiplier-accelerator model [23], the IS-LM/AD-AS model [13,59], Dale Jorgenson's investment model [51], William Nordhaus' [34] DICE climate change model [4,5], and basic micro economic supply and demand mechanisms [24]. Low's improvement of Hicks' model is particularly interesting because it results in a model that closely resembles Bill Phillips' [40] multiplier-accelerator model. Senge and Fiddaman's contributions are also very interesting because they demonstrate how the original economic models are special cases of their more general system dynamics formulations.

Creating Economic Dynamics Models from Scratch

Although translating well known economic models into a system dynamics format can arguably make them easier to understand and use, system dynamicists believe that the "proper" way to model an economic system that is experiencing a problem is to do so from scratch while following good system dynamics modeling practice. Unlike orthodox economists who generally follow a deductive, logical positivist approach to modeling, system dynamicists follow an inductive pattern modeling or case study process. More specifically, a system dynamicist approaches an economic problem like a detective who is iteratively piecing together an explanation at a crime scene. All types of data that are deemed relevant to the problem are considered including numerical, written, and mental information. The system dynamicist is guided in the pattern modeling process by the perceived facts of the case, as well as by real typologies (termed "generic structures" in system dynamics)



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 4
Simple two sector Keynesian cross model in a system dynamics format



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 5
Improved simple two sector Keynesian cross model

and principles of systems. Real typologies are commonalities that have been found to exist in different pattern models and principles of systems are commonalities that have been found to exist in different real typologies. Paich [36] discusses generic structures at length and Forrester [8] lays out a set of principles of systems.

Examples of a real typologies in economics include Forrester's [9] *Urban Dynamics* model, which can reproduce the behavior of many different cities when properly parametrized for those cities, and Homer's [22] model of the diffusion of new medical technologies into the market place, which can explain the behavior of a wide variety of medical technologies when properly parametrized for those technologies. Examples of fundamental principles of systems include the principle of accumulation, which states that the dynamic behavior of any system is due to flows accumulating in stocks, and the notion of stocks and flows being components of feedback loops. The parallels for these principles in economics can be found in modern Post Keynesian economics, in which modelers try to build "stock-flow consistent models," and in institutional economics, in which the principle of "circular and cumulative causation" is deemed to be a fundamental cause of economic dynamics. Radzicki [41,43,45] lays out the case for the parallels that exist between methodological concepts in system dynamics and methodological concepts in various schools of economic thought.

The economic models that have been historically created from scratch by following the system dynamics paradigm have tended to be fairly large in scale. Forrester's [12] national economic model is a classic example, as are the macroeconomic models created by Sterman [53], the Millennium Institute [31], Radzicki [45], Wheat [59], and Yamaguchi [60]. Dangerfield [3] has developed a model of Sarawak (E. Malaysia) to analyze and plan for economic transition from a production economy to a knowledge-based one. With the exception of Radzicki [45], whose model is based on ideas from Post Keynesian and institutional economics, these models, by and large, embody orthodox economic relationships.

Model Validity

When a system dynamics model of an economic system that is experiencing a problem is built from scratch, the modeling process is typically quite different from that which is undertaken in traditional economics. As such, the question is raised as to whether or not an original system dynamics model is in any sense "valid".

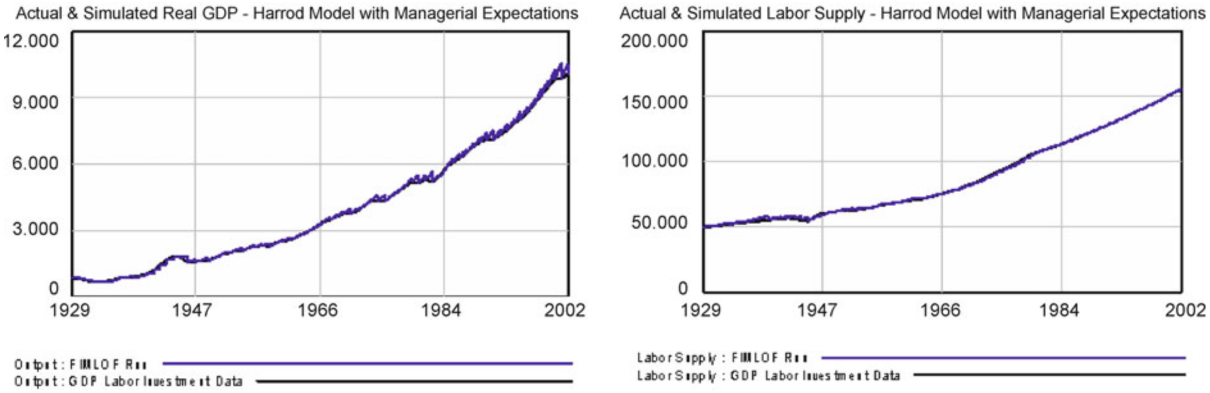
System dynamicists follow a "pattern modeling" approach [41] and do not believe that models should be judged in a binary fashion as either "valid" or "invalid". Rather, they argue that confidence in models can be generated along multiple dimensions. More specifically, system dynamicists such as Peterson [38], Forrester and Senge [16] and Barlas [1] have developed a comprehensive series of tests that can be applied to a model's structure and behavior and they argue that the more tests a model can pass, the more confidence a model builder or user should place in its results. Even more fundamentally, however, Forrester [13] has argued that the real value generated through the use of system dynamics comes, not from any particular model, but from the modeling *process* itself. In other words, it is through the iterative *process* of model conceptualization, creation, simulation, and revision that true learning and insight are generated, and *not* through interaction with the resulting model.

Another issue that lies under the umbrella of model validity involves fitting models to time series data so that parameters can be estimated and confidence in model results can be raised. In orthodox economics, of course, econometric modeling is almost universally employed when doing empirical research. Orthodox economic theory dictates the structure of the econometric model and powerful statistical techniques are used to tease out parameter values from numerical data.

System dynamicists, on the other hand, have traditionally argued that it is not necessary to tightly fit models to time series data for the purposes of parameter estimation and confidence building. This is because:

1. the battery of tests that are used to build confidence in system dynamics models go well beyond basic econometric analysis;
2. the particular (measured) time path that an actual economic system happened to take is merely one of an infinite number of paths that it could have taken and is a result of the particular stream of random shocks that happened to be historically processed by its structure. As such, it is more important for a model to mimic the basic character of the data, rather than fit it point-by-point [14];
3. utilizing the pattern modeling/case study approach enables the modeler to obtain parameter values via observation below the level of aggregation in the model, rather than via statistical analysis [18];
4. the result of a system dynamics modeling intervention is typically a set of policies that improve system performance and increase system robustness. Such policies are usually feedback-based rules (i. e., changes to institutional structure) that do not require the accurate point prediction of system variables.

Although the arguments against the need to fit models to time series data are well known in system dynamics, many system dynamicists feel that it is still a worthwhile activity because it adds credibility to a modeling study. Moreover, in modern times, advances in software technology have made this process relatively easy and inexpensive. Although several techniques for estimating the pa-



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 6
Fit of the Harrod growth model to US macroeconomic data for the years 1929–2002

rameters of a system dynamics model from numerical data have been devised, perhaps the most interesting is David Peterson's [38,39] Full Information Maximum Likelihood with Optimal Filtering (FIMLOF). Figure 5 presents a run from the Harrod growth model, to which an adaptive expectations structure has been added, after it has been fit via FIMLOF to real GDP and labor supply data for the United States economy for the years 1929–2002. The fit is excellent and the estimated parameter values are consistent with those from more traditional econometric studies. See Radzicki [44] for a detailed description of the model and its parameter estimates.

Controversies

Since system dynamics modeling is undertaken in a way that is significantly different from traditional economic modeling, it should come as no surprise that many economists have been extremely critical of some system dynamics models of economic systems. For example, Forrester's [9] *Urban Dynamics* and [10] *World Dynamics* models have come under severe attack by economists, as has (to a lesser degree) his national economic model. On the other hand, the first paper in the field of system dynamics is Forrester [6], which is essentially a critique of traditional economic modeling.

Greenberger et al. [19] present a nice overview of the controversies surrounding the *Urban Dynamics* and *World Dynamics* models. Forrester and his colleagues' replies to criticisms of the *Urban Dynamics* model are contained in Mass [25] and Schroeder et al. [50].

One of the harshest critics of the *World Dynamics* (WORLD2) model has been Nordhaus [33]. Nordhaus [35] has also very critical of the well known follow-up study to *World Dynamics* known as *The Limits to Growth* [28]. Meadows et al. [29,30] contain updates to the original *Limits to Growth* (WORLD3) model, as well as replies to the world modeling critics.

Forrester [12] presents a nice overview of his national economic model, and the critiques by Stolwijk [57] and Zellner [61] are typical of the attitude of the professional economists toward macroeconomic modeling that is undertaken by following the traditional system dynamics paradigm. The criticism of Forrester's national economic model by the economics profession has probably been less severe, relative to the criticisms of the *Urban Dynamics* and world models, because most of its details are still largely unpublished at the time of this writing.

Another interesting and timely example of the sort of controversy surrounding system dynamics modeling in economics is provided by Sterman and Richardson [56]. In this paper they present a technique for testing whether Hubbert's lifecycle method or the geologic analogy method yields superior estimates of the ultimately recoverable amount of petroleum resources. This study was motivated by a disagreement with a traditionally trained economist over the proper way to conceptualize this issue. Sterman and Richardson devised a clever synthetic data experiment in which a system dynamics model serves as the "real world" with a known ultimately recoverable amount of oil. Hubbert's method and the geologic analogy method are then programmed into the model so they can "watch" the data being generated by the "real world" and provide dynamic estimates of the "known" ultimately recoverable stock of oil. The results showed that Hubbert's method was quite accurate, although it had a tendency to somewhat underestimate the ultimately recoverable amount of oil, while the geologic analogy method tended to overshoot the resource base quite substantially.

Future Directions

Historically, system dynamicists who have engaged in economic modeling have almost never been trained as professional economists. As such, they have had the advantage of being able to think about economic problems differently from those who have been trained along traditional lines, but have also suffered the cost of being seen as "amateurs" or "boy economists" [41] by members of the economics profession. The good news is that there are currently several schools of economic thought, populated by professional economists, in which system dynamics fits quite harmoniously. These include Post Keynesian economics, institutional economics, ecological economics, and behavioral economics. Historically, the economists in these schools have rejected many of the tenets of traditional economics, including most of its formal modeling methods, yet have failed to embrace alternative modeling techniques because they were all seen as inadequate for representing the concepts they felt were important. However in the modern era, with computers having become ubiquitous and simulation having become in some sense routine, system dynamics is increasingly being accepted as an appropriate tool for use in these schools of economic thought. The future of economics and system dynamics will most probably be defined by the economists who work within these schools of thought, as well as by their students. The diffusion of system dynamics models of economic systems through their translation into user-friendly interactive "learning environments" that are available over the world wide web will most likely also be of great importance (see [24,59]).

Bibliography

Primary Literature

1. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Operat Res 42(1):59–87
2. Barr N (1988) The Phillips Machine. LSE Q Winter 2(4):305–337
3. Dangerfield BC (2007) System dynamics advances strategic economic transition planning in a developing nation. In: Qudrat-Ullah H, Spector M, Davidsen P (eds) Complex decision-making: Theory & practice. Springer, New York, pp 185–209
4. Fiddaman T (1997) Feedback complexity in integrated climate-economy models. Ph D Dissertation, Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
5. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
6. Forrester J (1957) Dynamic models of economic systems and industrial organizations. System Dynamics Group Memo D-0. Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
7. Forrester J (1961) Industrial dynamics. Pegasus Communications, Inc., Waltham
8. Forrester J (1968) Principles of systems. MIT Press, Cambridge
9. Forrester J (1969) Urban dynamics. Pegasus Communications, Inc., Waltham
10. Forrester J (1971) World dynamics. Pegasus Communications, Inc., Waltham
11. Forrester J (1975) Counterintuitive behavior of social systems. In: Forrester J (ed) Collected papers of Jay W Forrester, Pegasus Communications, Inc., Waltham, pp 211–244
12. Forrester J (1980) Information sources for modeling the national economy. J Am Statist Assoc 75(371):555–567
13. Forrester J (1985) 'The' model versus a modeling 'process'. Syst Dyn Rev 1(1 and 2):133–134
14. Forrester J (2003) Economic theory for the new millennium. In: Eberlein R, Diker V, Langer R, Rowe J (eds) Proceedings of the Twenty-First Annual Conference of the System Dynamics Society. Available from http://www.systemdynamics.org/
15. Forrester J, Low G, Mass N (1974) The debate on world dynamics: A response to Nordhaus. Policy Sci 5:169–190
16. Forrester J, Senge P (1980) Tests for building confidence in system dynamics models. In: Legasto Jr AA, Forrester JW, Lyneis JM (eds) TIMS Studies in the Management Sciences: System Dynamics, vol 14. North Holland Publishing Company, Amsterdam, pp 209–228
17. Forrester N (1982) A dynamic synthesis of basic macroeconomic theory: Implications for stabilization policy analysis, Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
18. Graham A (1980) Parameter estimation in system dynamics modeling. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 143–161
19. Greenberger M, Crenson M, Crissey B (1976) Models in the policy process: Public decision making in the computer era, Russell Sage Foundation, New York
20. Heilbroner R (1980) The worldly philosophers, 5th edn. Simon & Schuster, New York
21. Hicks J (1950) A contribution to the theory of the trade cycle. Oxford University Press, London
22. Homer J (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Change 31(3):197–218
23. Low G (1980) The multiplier-accelerator model of business cycles interpreted from a system dynamics perspective. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 76–94
24. Mashayekhi A, Vakili K, Foroughi H, Hadavandi M (2006) Supply demand world: An interactive learning environment for teaching microeconomics. In: Grosler A, Rouwette A, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fourth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
25. Mass N (1974) Readings in urban dynamics, vol I. Wright-Allen Press, Cambridge

26. Mass N (1975) Economic cycles: An analysis of the underlying causes. MIT Press, Cambridge
27. Meadows D (1970) Dynamics of commodity production cycles. Massachusetts MIT Press, Cambridge
28. Meadows D, Meadows D, Randers J, Behrens III W (1972) The limits to growth: A report for the Club of Rome's project on the predicament of mankind. Universe Books, New York
29. Meadows D, Meadows D, Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green Publishing, White River Junction
30. Meadows D, Meadows D, Randers J (2002) Limits to growth: The 30-year update. Chelsea Green Publishing, White River Junction
31. Millennium Institute (2007) Introduction and Purpose of Threshold 21, http://www.millennium-institute.org/resources/elibrary/papers/T21Overview.pdf
32. Moxnes E (1992) Positive feedback economics and the competition between 'hard' and 'soft' energy supplies. J Sci Ind Res 51(March):257–265
33. Nordhaus W (1972) World dynamics: Measurement without data. Econom J 83(332):1156–1183
34. Nordhaus W (1992) The "DICE" model: Background and structure of a dynamic integrated climate-economy model of the economics of global warming. Cowles Foundation for Research in Economics at Yale University, Discussion Paper No. 1009
35. Nordhaus W (1992) Lethal model 2: The limits to growth revisited. Brookings Pap Econo Activity 2:1–59
36. Paich M (1985) Generic structures. Syst Dyn Rev 1(1 and 2):126–132
37. Paich M (1994) Managing the global commons. MIT Press, Cambridge
38. Peterson D (1975) Hypothesis, estimation, and validation of dynamic social models – energy demand modeling. Ph D Dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
39. Peterson D (1980) Statistical tools for system dynamics. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 226–245
40. Phillips W (1954) Stabilization policy in a closed economy. Econom J 64(254):290–323
41. Radzicki M (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147
42. Radzicki M (1997) Introduction to system dynamics. Free web-based system dynamics tutorial. Available at http://www.systemdynamics.org/DL-IntroSysDyn/index.html
43. Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. J Econom Issues 37(1):133–173
44. Radzicki M (2004) Expectation formation and parameter estimation in nonergodic systems: the system dynamics approach to post Keynesian-institutional economics. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the Twenty-Second International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
45. Radzicki M (2007) Institutional economics, post keynesian economics, and system dynamics: Three strands of a heterodox economics braid. In: Harvey JT, Garnett Jr. RF (eds) The future of heterodox economics. University of Michigan Press, Ann Arbor
46. Radzicki M, Sterman J (1994) Evolutionary economics and system dynamics. In: Englund R (ed) Evolutionary concepts in contemporary economics. University of Michigan Press, Ann Arbor, pp 61–89
47. Richardson G, Pugh A (1981) Introduction to system dynamics modeling with DYNAMO. Pegasus Communications, Inc., Waltham
48. Richmond B (1985) Conversing with a classic thinker: An illustration from economics. Users Guide to STELLA, Chapt 7. High Performance Systems, Inc., Lyme, New Hampshire, pp 75–94
49. Ryzhenkov A (2007) Controlling employment, profitability and proved non-renewable reserves in a theoretical model of the US economy. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
50. Schroeder W, Sweeney R, Alfeld L (1975) Readings in Urban Dynamics, vol II. Wright-Allen Press, Cambridge
51. Senge P (1980) A system dynamics approach to investment function formulation and testing. Soc-Econ Plan Sci 14:269–280
52. Solow R (1956) A contribution to the theory of economic growth. Q J Econom 70:65–94
53. Sterman J (1981) The energy transition and the economy: A system dynamics approach. Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available at http://www.systemdynamics.org/
54. Sterman J (1985) A behavioral model of the economic long wave. J Econom Behav Organ 6:17–53
55. Sterman J (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin-McGraw-Hill, New York
56. Sterman J, Richardson G (1985) An experiment to evaluate methods for estimating fossil fuel resources. J Forecast 4(2):197–226
57. Stolwijk J (1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):569–572
58. Weber L (2007) Understanding recent developments in growth theory. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
59. Wheat D (2007) The feedback method of teaching macroeconomics: Is it effective? Syst Dyn Rev 23(4):391–413
60. Yamaguchi K (2007) Balance of payments and foreign exchange dynamics – S D Macroeconomic Modeling (4). In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
61. Zellner A(1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):567–569

Books and Reviews

Alfeld L, Graham A (1976) Introduction to urban dynamics. Wright-Allen Press, Cambridge

Lyneis J (1980) Corporate planning and policy design: A system dynamics approach. PA Consulting, Cambridge

Meadows D, Meadows D (1973) Toward global equilibrium: Collected papers. Wright-Allen Press, Cambridge

Meadows D, Behrens III W, Meadows D, Naill R, Randers J, Zahn E (1974) Dynamics of growth in a finite world. Wright-Allen Press, Cambridge

Meadows D, Robinson J (1985) The electronic oracle: Computer models and social decisions. Wiley, New York

Richardson G (1999) Feedback thought in social science and systems theory. Pegasus Communications, Inc., Waltham

Sterman J (1988) A skeptic's guide to computer models. In: Grant L (ed) Foresight and National Decisions. University Press of America, Lanham, pp 133–169, Revised and Reprinted in: Barney G, Kreutzer W, Garrett M (1991) Managing a Nation. Westview Press, Boulder, pp 209–230

System Dynamics and Organizational Learning

KAMBIZ MAANI
Chair in Systems Thinking and Practice,
The University of Queensland, Brisbane, Australia

Article Outline

Glossary

Stock In system dynamics stock is a concept representing accumulation and the state of a variable, such as, assets, inventory, capacity, reputation, morale etc. Stock can be measured at any point of time. In mathematical terms, stock is the sum over time (integral) of one or more flows.

Flow Flow or rate represents change or movement in a stock such as, buying assets, building inventories, adding capacity, losing reputation or morale, etc. Flow is measured as "per unit of time" like hiring rate (employees hired per year, production rate (units made per day), or rainfall (inches of rain per month).

Causal loops Causal loops (model) are visual maps that connect a group of variables with known or hypothesized cause and effect relationships. A causal loop can be open or closed. Causal loops can be used for complex problem solving/decision making, consensus building, conflict resolution, priority setting and group learning.

Feedback In a cause and effect chain (system), feedback is a signal from the effect/s to cause/s as to its/their influence on downstream effect/s. Feedback can be information, decision or action. For example, if X causes or changes Y, Y in turn could influence or change X directly or through other intervening variables. This creates a *closed* "causal loop" with either a positive or amplifying feedback (Reinforcing – R) or a negative feedback with damping, counteracting or (Balancing – B) effect.

Delay Cause and effect relationships are often not close in time or space. The lapse time between a cause and its effect is called a systems delay or simply delay. Because some delays in physical, natural and social systems are rather long they mask the underlying or earlier causes when effects become evident. This provides confusion and unintended consequences, especially in social systems, such as economics, education, immigration, judicial systems, etc.

Reference Mode Reference mode is the actual/observed pattern of a key variable of interest to decision makers or policy analysts. It represents the actual behavior of a variable over time which is used to compare with the simulated pattern of the same variable generated by a simulation model to validate the accuracy of the model.

Simulation A computer tool and methodology for modeling complex situations and challenging problems where mathematical tools fail to operate.

Microworld Microworlds are simulation models of real systems such as a firm, a hospital, a market, or a production system. They provide a "virtual" world where decision makers can test and experiment their policies and strategies in a laboratory environment before implementation. Microworlds are constructed using system dynamic software with user friendly interfaces.

Leverage Leverage refers to decisions and actions for change and intervention which have the highest likelihood of lasting and sustainable outcomes. Leverage decisions are best reached by open discussion after the group develops a deep understanding of system dynamics through a causal loop or stock & flow modeling process.

Systems thinking Systems thinking is a paradigm for viewing reality based on the primacy of the whole and relationships. It is one of the key capabilities (disciplines) for organizational learning [30]. Systems Thinking consists of a series of conceptual and modeling tools such as behavior over time, causal loop diagrams and systems archetypes. These tools reveal cause and effect dynamics over time and assist understanding of complex, non-linear, and counter-intuitive behaviors in all systems – physical, natural and social.

Definition of the Subject

System dynamics (SD) is "a methodology for studying and managing complex feedback systems... While the word system has been applied to all sorts of situations, feedback is the differentiating descriptor here. Feedback refers to the situation of X affecting Y and Y in turn affecting X perhaps through a chain of causes and effects... Only the study of the whole system as a feedback system will lead to correct results." [36]

Sterman ([35], p 4) defines System Dynamics as "a method to enhance learning in complex systems". "System dynamics is fundamentally interdisciplinary… It is grounded in the theory of nonlinear dynamics and feedback control developed in mathematics, physics, and engineering. Because we apply these tools to the behavior of human as well as physical and technical systems, system dynamics draws on cognitive and social psychology, economics, and other social sciences."

Wolstenholme's [40] offers the following description for system dynamics and its scope:

> A rigorous way to help thinking, visualizing, sharing, and communication of the future evolution of complex organizations and issues over time; for the purpose of solving problems and creating more robust designs, which minimize the likelihood of unpleasant surprises and unintended consequences; by creating operational maps and simulation models which externalize mental models and capture the interrelationships of physical and behavioral processes, organizational boundaries, policies, information feedback and time delays; and by using these architectures to test the holistic outcomes of alternative plans and ideas; within a framework which respects and fosters the needs and values of awareness, openness, responsibility and equality of individuals and teams.

Organizational Learning

Organizational learning is the ability of organizations to enhance their collective capacity to learn and to act, harmoniously. According to Senge [30] "Real learning gets to the heart of what it means to be human. Through learning we re-create ourselves. Through learning we become able to do something we never were able to do. Through learning we re-perceive the world and our relationship to it. Through learning we extend our capacity to create, to be part of the generative process of life. There is within each of us a deep hunger for this type of learning." Organizational learning extends this learning to the organization and its members.

Introduction

History of System Dynamics

(This section is due to US Department of Energy website) "System dynamics was created during the mid-1950s by Professor Jay W. Forrester of the Massachusetts Institute of Technology. Forrester arrived at MIT in 1939 for graduate study in electrical engineering. His first research assistantship put him under the tutelage of Professor Gordon Brown, the founder of MIT's Servomechanism Laboratory. Members of the MIT Servomechanism Laboratory, at the time, conducted pioneering research in feedback control mechanisms for military equipment. Forrester's work for the Laboratory included traveling to the Pacific Theatre during World War II to repair a hydraulically controlled radar system installed aboard the aircraft carrier Lexington. The Lexington was torpedoed while Forrester was on board, but not sunk.

At the end of World War II, Jay Forrester turned his attention to the creation of an aircraft flight simulator for the US Navy. The design of the simulator was cast around the idea, untested at the time, of a digital computer. As the brainstorming surrounding the digital aircraft simulator proceeded, however, it became apparent that a better application of the emerging technology was the testing of computerized combat information systems. In 1947, the MIT Digital Computer Laboratory was founded and placed under the direction of Jay Forrester. The Laboratory's first task was the creation of WHIRLWIND I, MIT's first general-purpose digital computer, and an environment for testing whether digital computers could be effectively used for the control of combat information systems. As part of the WHIRLWIND I project, Forrester invented and patented coincident-current random-access magnetic computer memory. This became the industry standard for computer memory for approximately twenty years. The WHIRLWIND I project also motivated Forrester to create the technology that first facilitated the practical digital control of machine tools.

After the WHIRLWIND I project, Forrester agreed to lead a division of MIT's Lincoln Laboratory in its efforts to create computers for the North American SAGE (Semi-Automatic Ground Environment) air defense system. The computers created by Forrester's team during the SAGE project were installed in the late 1950s, remained in service for approximately twenty-five years, and had a remarkable "up time" of 99.8%.

Forrester's seminal book Industrial Dynamics [11] "is still a significant statement of philosophy and methodology in the field. Since its publication, the span of applications has grown extensively and now encompasses work in

- corporate planning and policy design
- public management and policy
- biological and medical modeling
- energy and the environment

System Dynamics and Organizational Learning, Table 1
The five phase process of systems thinking and modeling (Source: [6])

Phases	Steps	
1	Problem structuring	Identify problems or issues of concern to management, Collect preliminary information and data
2	Causal loop modeling	Identify main variables, Prepare behavior over time graphs (reference mode), Develop causal loop diagram (influence diagram), Analyze loop behavior over time and identify loop types, Identify system archetypes, Identify key leverage points, Develop intervention strategies
3	**System dynamic modeling**	Develop a systems map or rich picture, Define variable types and construct stock-flow diagrams, Collect detailed information and data, Develop a simulation model, Simulate steady -state/stability conditions, Reproduce reference mode behavior (base case), Validate the model, Perform sensitivity analysis, Design and analyze policies, Develop and test strategies
4	Scenario planning and modeling	Plan general scope of scenarios, Identify key drivers of change and keynote uncertainties, Construct forced and learning scenarios, Simulate scenarios with the model, Evaluate robustness of the policies and strategies
5	Implementation and organizational learning	Prepare a report and presentation to management team, Communicate results and insights of proposed intervention to stakeholders, Develop a microworld and learning lab based on the simulation model, Use learning lab to examine mental models and facilitate

- theory development in the natural and social sciences
- dynamic decision making
- complex nonlinear dynamics” [36]

Systems Thinking and Modeling Methodology

System Dynamics is one of the five phases of systems thinking and modeling intervention methodology [6,21]. These distinct but related phases are as follows:

1. Problem structuring;
2. Causal loop modeling;
3. System dynamics modeling;
4. Scenario planning and modeling;
5. Implementation and organizational learning (learning lab).

These phases follow a process, each involving a number of steps, as outlined in Table 1. This process does not require all phases to be undertaken, nor does each phase require all the steps listed. Which phases and steps are included in a particular project or intervention depends on the issues or problems that have generated the systems enquiry and the degree of effort that the organization is prepared to commit to.

System Dynamics Modeling

This phase follows the causal modeling phase. Although it is possible to go into this phase directly after problem structuring, performing the causal modeling phase first will enhance the conceptual rigor and learning power of the systems approach. The completeness and wider insights of systems thinking is generally absent from other simulation modeling approaches, where causal modeling does not play a part. The following steps are generally followed in the system dynamics modeling phase.

1. Develop a high-level map or systems diagram showing the main sectors of a potential simulation model, or a ‘rich picture’ of the main variables and issues involved in the system of interest.
2. Define variable types (e. g. stocks, flows, converters, etc.) and construct stock flow diagrams for different sectors of the model.

3. Collect detailed, relevant data including media reports, historical and statistical records, policy documents, previous studies, and stakeholder interviews.
4. Construct a computer simulation model based on the causal loop diagrams or stock-flow diagrams. Identify the initial values for the stocks (levels), parameter values for the relationships, and the structural relationships between the variables using constants, graphical relationships and mathematical functions where appropriate. This stage involves using specialized computer packages like STELLA, *ithink*, VENSIM, POWERSIM, DYSMAP, COSMIC and Consideo.
5. Simulate the model over time. Select the initial value for the beginning of the simulation run, specify the unit of time for the simulation (e. g. hour, day, week, month, year, etc.). Select the simulation interval (DT) (e. g. 0.25, 0.5, 1.0) and the time horizon for the simulation run (i. e. the length of the simulation). Simulate model stability by generating steady state conditions.
6. Produce graphical and tabular output for the base case of the model. This can be produced using any of the computer packages mentioned above. Compare model behavior with historical trends or hypothesized reference modes (behavior over time charts).
7. Verify model equations, parameters and boundaries, and validate the model's behavior over time. Carefully inspect the graphical and tabular output generated by the model.
8. Perform sensitivity tests to gauge the sensitivity of model parameters and initial values. Identify areas of greatest improvement (key leverage points) in the system.
9. Design and test policies with the model to address the issues of concern to management and to look for system improvement.
10. Develop and test strategies (i. e. combinations of functional policies, for example operations, marketing, finance, human resources, etc.).

Organizational Learning

(This section is adapted from [21])
Peter Senge, who popularized the concept through his seminal book: The Fifth Discipline [30], describes a learning organization as one 'which is continually expanding its ability to create its future'. He identifies five core capabilities (disciplines) of the learning organization that are derived from three "higher orientations": creative orientation; generative conversation; and systems perspective. "The reality each of us sees and understands depend on what we believe is there. By learning the principles of the five disciplines, teams begin to understand how they can think and inquire that reality, so that they can collaborate in discussions and in working together create the results that matter [to them]."

As Fig. 1 shows the learning organization capabilities are dynamically interrelated, and collectively they lead to organizational learning.

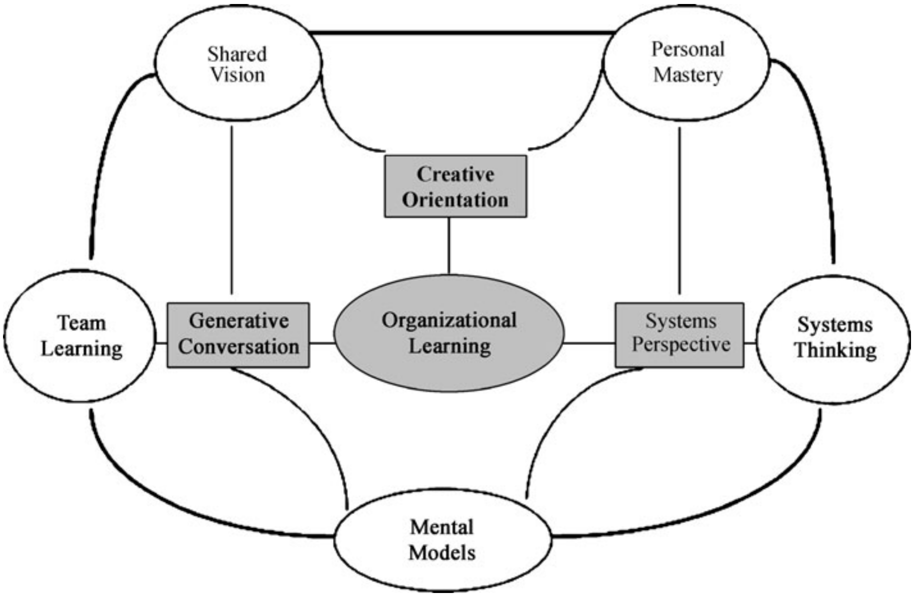
Senge maintains that *Creative orientation* is the source of a genuine desire to excel. It is the source of an intrinsic motivation and drive to achieve. It relinquishes personal gains in favor of the common good. *Generative conversation* refers to a deep and meaningful dialog to create unity of thought and action. *Systems perspective* is the ability to see things holistically by understanding the interconnectedness of the parts. The foregoing elements give rise to the five core capabilities of learning organizations, namely: personal mastery; shared vision; mental models; team learning and dialog; and systems thinking. These five disciplines are described below. Figure 1 below shows the core capabilities and their relationships.

Personal Mastery

Senge [30] describes that personal mastery is the cornerstone and 'spiritual' foundation of the learning organization. It is born out of a creative orientation and systemic perspective. Personal mastery instils a genuine desire to do well and to serve a noble purpose. People exhibiting high levels of personal mastery focus "on the desired result itself, not the process or the means they assume necessary to achieve that result" [30]. These people can "successfully focus on their ultimate intrinsic desires, not on secondary goals. This is a cornerstone of Personal Mastery". Personal mastery also requires a commitment to truth, which means to continually challenge "theories of why things are the way they are". Without committing to the truth, people all too quickly revert to old communication routines which can distort reality and prevent them from knowing where they really stand.

Shared Vision

It is commonly assumed that in contemporary organizations senior management can develop a vision which employees will follow with genuine commitment. This is a fallacy. Simply promoting a 'vision statement' could result in a sense of apathy, complacency and resentment. Instead, there needs to be a genuine endeavor to understand what people will commit to. The overriding vision of the group must build on the personal visions of its members. Shared



System Dynamics and Organizational Learning, Figure 1
The core capabilities of a learning organization (Source: [19])

vision should align diverse views and feelings into a unified focus.

This is emphasized by Arie de Geus [9] when he describes what makes a truly extraordinary organization. "The feeling of belonging to an organization and identifying with its achievements is often dismissed as soft. But case histories repeatedly show that a sense of community is essential for long term survival". For example, when Apple Corporation challenged IBM, it was in its 'adolescent' years, characterized by creativity, confidence and even defiance. This is similar to the spirit in Team New Zealand when it competed against the bigger-budget syndicates! Within these organizations there is a real passion for the outcome; a common vision for success [19].

Creating a shared vision is the most fundamental job of a leader [26]. By creating a vision, the leader provides a vehicle for people to develop commitment, a common goal around which people can rally, and a way for people to *feel* successful. The leader must appeal to people's emotions if they are to be energized towards achieving the goal. Emotional acceptance of, and belief in, a vision is far more powerful in energizing team members than is intellectual recognition that the vision is simply a 'good idea'. One of the most powerful ways of communicating a vision is through a leader's personal example and actions, demonstrating behavior that symbolizes and furthers that vision.

Mental Model and Leadership

Mental models reflect beliefs, assumptions and feelings that shape one's world views and actions. They are formed through family, education, professional and social learning based, on the most part, on cultural and social norms. Mental models, however, can be altered and aligned.

Organizations are often constrained by deep-seated belief systems, resulting in preconceived ideas on how things ought to perform. Goodstein and Burke ([14] p. 10), pioneers in the field of social psychology of organizations, observed that 'the first step in any change process is to unfreeze the present patterns of behavior as a way of managing resistance to change'. The leader has a pivotal role in dismantling negative mental models and shaping new ones.

In order to get people to engage in open discussions of issues that affect the organization, a leader must appeal to their emotions and must get beyond the superficial level of communication. In the 1970s Shell Oil undertook major changes in its leadership approach and communications style. According to a manager at Shell, "When I tried to talk personally about an issue rather than say 'here's the answer', it was powerful. It caused me to engage in dialog with others that resulted in mutual learning on all sides" ([7] p. 71).

The leader is a 'designer', and part of that role is designing the governing ideas of purpose and core values by which people will live [30,31]. In this role, the leader must propose and model the manner in which the group has to operate internally. This provides ample opportunities for leaders to examine their deeply held assumptions about the task, the means to accomplish it, the uniqueness of the people and the kinds of relationship that should be fostered among the people. Only after people have observed and *experienced* the organizational values in practice would these values become the basis for prolonged group behavior. These values should be manifested first and should be most visible in the leader's own behavior.

Leadership, especially in knowledge-based organizations, must be distributed and shared to a far greater extent than it was in the past. For example, in the Chicago Bulls basketball team, Michael Jordan changed his role: it became not only that of an individually brilliant player but *also* that of a leader whose job it was to raise the level of play of other team members. After this transition, the Bulls began their record run of championship seasons [7].

Team Learning and Dialog

The word 'dialog' comes from the Greek words *dia* and *logos*. It implies that when people engage in dialog, the meaning *moves through* them – Thus, it enables them to 'see through words' [16]. Dialog is an essential requirement for organizational learning. It results from generative conversation, shared vision, and transparent mental models. Dialog creates a deep sense of listening and suspending one's own views. Feedback is an integral aspect of dialog.

Communication routines in organizations are generally anti-learning and promote mediocrity. They include 'defensive routines' [2] – statement that can stifle dialog and innovative thinking. Exposing and unlearning such routines, and understanding the powerful detrimental impact they have on learning, are serious challenges many organizations face if they are to create effective learning environments.

Many leaders are charismatic and are highly eloquent when it comes to presenting their ideas; that's often why they get to the top of the organization. However, many appear to lack the ability to extract the very best from employees in a non-threatening manner. Without this ability, leaders may miss many good ideas, or might act on many bad ones.

In a group context, encouragement from the leader and mutual encouragement among group members is essential. Furthermore, personal differences must be put aside in order for effective dialog to ensue.

How Organizations Learn

(This section is edited from Wikipedia: http://en.wikipedia.org/wiki/Organizational_learning)
"Argyris and Schon were the first to propose concepts and models that facilitate organizational learning, the following literatures have followed in the tradition of their work:

- March and Olsen [23] attempt to link up individual and organizational learning. In their model, individual beliefs lead to individual action, which in turn may lead to an organizational action and a response from the environment which may induce improved individual beliefs and the cycle then repeats over and over. Learning occurs as better beliefs produce better actions.
- Argyris and Schon [3] distinguish between single-loop and double-loop learning, related to Gregory Bateson's concepts of first and second order learning. In single-loop learning, individuals, groups, or organizations modify their actions according to the difference between expected and obtained outcomes. In double-loop learning, the entities (individuals, groups or organization) question the values, assumptions and policies that led to the actions in the first place; if they are able to view and modify those, then second-order or double-loop learning has taken place. Double loop learning is the learning about single-loop learning.
- Kim [17], as well, in an article titled "The link between individual and organizational learning", integrates Argyris, March and Olsen and another model by Kofman into a single comprehensive model; Further, he analyzes all the possible breakdowns in the information flows in the model, leading to failures in organizational learning; For instance, what happens if an individual action is rejected by the organization for political or other reasons and therefore no organizational action takes place?
- Nonaka and Takeuchi [27] developed a four stage spiral model of organizational learning. They started by differentiating Polanyi's concept of "tacit knowledge" from "explicit knowledge" and describe a process of alternating between the two. Tacit knowledge is personal, context specific, subjective knowledge, whereas explicit knowledge is codified, systematic, formal, and easy to communicate. The tacit knowledge of key personnel within the organization can be made explicit, codified in manuals, and incorporated into new products and processes. This process they called "externalization". The reverse process (from explicit to implicit) they call "internalization" because it involves employees internalizing an organization's formal rules, procedures, and other forms of explicit knowledge. They

also use the term "socialization" to denote the sharing of tacit knowledge, and the term "combination" to denote the dissemination of codified knowledge. According to this model, knowledge creation and organizational learning take a path of socialization, externalization, combination, internalization, socialization, externalization, combination… etc. in an infinite spiral.

- Flood [10] discusses the concept of organizational learning from Peter Senge and the origins of the theory from Argyris and Schon. The author aims to "re-think" Senge's *The Fifth Discipline* through systems theory. The author develops the concepts by integrating them with key theorists such as Bertalanffy, Churchman, Beer, Checkland and Ackoff. Conceptualizing organizational learning in terms of structure, process, meaning, ideology and knowledge, the author provides insights into Senge within the context of the philosophy of science and the way in which systems theorists were influenced by twentieth-century advances from the classical assumptions of science.
- Nick Bontis et al. [4] empirically tested a model of organizational learning that encompassed both stocks and flows of knowledge across three levels of analysis: individual, team and organization. Results showed a negative and statistically significant relationship between the misalignment of stocks and flows and organizational performance.
- Imants [15] provides theory development for organizational learning in schools within the context of teachers' professional communities as learning communities, which is compared and contrasted to teaching communities of practice. Detailed with an analysis of the paradoxes for organizational learning in schools, two mechanisms for professional development and organizational learning, (1) steering information about teaching and learning and (2) encouraging interaction among teachers and workers, are defined as critical for effective organizational learning.
- Common [8] discusses the concept of organizational learning in a political environment to improve public policy-making. The author details the initial uncontroversial reception of organizational learning in the public sector and the development of the concept with the learning organization. Definitional problems in applying the concept to public policy are addressed, noting research in UK local government that concludes on the obstacles for organizational learning in the public sector: (1) overemphasis of the individual, (2) resistance to change and politics, (3) social learning is self-limiting, i. e. individualism, and (4) political"blame culture". The concepts of *policy learning* and *policy transfer* are then defined with detail on the conditions for realizing organizational learning in the public sector."

Modeling for Organizational Learning

In general, the *process* of model building can be an effective conduit for collective learning. System Dynamics modeling, in particular, can be used to enhance organizational learning [35] through rapid feedback and experimentation and its facility to test assumptions and mental models. As we have discussed, dealing effectively with mental models is one of the core competencies for organizational learning

Ackoff [1] likens complex problems to "messes". "Messy problems are defined as situations in which there are large differences of opinion about the problem or even on the question of whether there is a problem. Messy situations make it difficult for a management team to reach agreement. System Dynamics modeling with groups known as Group Model Building (GMB) is a powerful tool for dealing with these. SD and GMB are especially effective in dealing with semi-structured and ill-structured decision situations."

GMB offers an opportunity to align and share piecemeal mental models and create the possibility of assimilating and integrating partial mental models into a holistic system description [38,39]. GMB and SD can help uncover 'illusions' that may occur due to the fact that the definition of a problem may be a socially constructed phenomenon that has not been put to test ([18] p. 84).

Learning Laboratory

(This section is adapted from [21], Chapter 6)
Learning laboratory is a setting as well as a process in which a group can learn together. The purpose of the learning lab is to enable managers to test their long held assumptions and to experiment and 'see' the consequences of their actions, policies and strategies. This often results in finding inconsistencies and the discovery of *unintended* consequences of actions and decisions, *before* they are implemented. System Dynamics models known as Microworlds or Management Flight Simulators (MFS) are the 'engine' behind the learning lab. "Just as an airline uses flight simulators to help pilots learn, system dynamics is, partly, a method for developing management flight simulators, often computer simulation models, to help us learn about dynamic complexity, understand the sources of policy resistance, and design more effective policies." ([35], p. 4)

A learning lab is distinct from so-called management games. In management games, the players are required to compete – design the 'best' strategy and 'beat' other players or teams. The competitive nature of management games often encourages aggressive and individualistic behavior with scant regard for group learning and gaining deep insights. The learning lab, in contrast, aims to enhance *learning*: To test individual and group mental models and to provide deeper understanding and insights into why systems behave the way they do. This will help the participants to test their theories and discover inconsistencies and 'blind spots' in policies and strategies *before* they are implemented.

A significant benefit of the learning lab stems from the process in which participants examine, reveal and test their mental models and those of their organization. The learning lab can also help participants

- To align strategic thinking with operational decisions;
- To connect short-term and long-term measures;
- To facilitate integration within and outside the organization;
- To undertake experimentation and learning;
- To balance competition with collaboration.

Managerial Practice Field

Team and teamwork are parts of the lexicons of numerous organizations today. Company after company has reorganized work around a variety of team concepts. From factories to hospitals, *titles* like 'manager' and 'supervisor' have been replaced by *roles* such as 'facilitator' and 'team leader'. Despite this level of attention to team and teamwork the expected benefits have been marginal at best.

But when we examine real teams, such as sporting teams, orchestras or ballet companies more closely, they all share one key characteristic. That is they *practice* a lot more than they 'perform'. Practice involves allowing time and space to experiment with new ways, try different approaches and most importantly, make mistakes without the fear of failure. In fact, making mistakes is indispensable to learning. One cannot learn from doing things right all the time! Yet a great deal of organizational energy and attention is devoted to the prevention and masking of mistakes.

But, what is the *practice field* for management teams? The fact is that the practice field is, by and large, absent from the managerial world. In other words, there is no time and no space for management to 'practice' in the true sense of the word – to experiment, make mistakes and learn together. In this era of restructuring and downsizing, lack of time is the greatest impediment to managerial and organizational learning. As a recent advertisement by IBM reads, "Innovative Thinking! We don't even have time for bad thinking". The pace in the modern work environment is so unrelenting that there is virtually no room for managers to slow down, to practice, to reflect and learn. The consequence of this lack of practice and learning space is grave, in that most organizations only achieve a small fraction of their potential – about 5%, according to Jay Forrester, the father of System Dynamics [13].

In order to fill this gap, the concept of learning laboratory has been developed to provide practice fields for managers. The learning lab allows learning to become an integral part of managerial work and helps learning to become institutionalized [17].

Aligning Mental Models Through the Learning Laboratory

Mental models are formed throughout one's life. Family, school, culture, religion, profession and social norms play important roles in this formation. Therefore, modifying one's mental model is not a small matter. The most effective way to check one's mental models is to *experience* alternative realities at first hand and see their implications with a new 'lens' [5].

There are rarely any opportunities in the course of a manager's daily work for him/her to engage in lengthy, drawn-out experimentation. Learning in a 'laboratory' setting is a viable and powerful alternative. Fortunately, advanced computers and sophisticated system dynamics software have enabled the creation of managerial learning labs where managers can experiment, test their theories and learn rapidly. Thus, learning labs can play a significant role in clarifying and changing mental models. Learning lab deals with mental models at three levels [33], as described below.

- *Mapping* mental models. This step begins at the conceptualization phase. Here, the learning lab participants articulate and clarify their assumptions, views, opinions, and biases regarding the issue at hand.
- *Challenging* mental models. The participants identify and discuss inconsistencies and contradictions in their assumptions. This step will begin at the conceptualizations phase and will continue to the experimentation phase.
- *Improving* mental models. Having conducted experimentation and testing, the participants reflect on the outcomes. This may cause them to alter, adjust, improve and harmonize their mental models.

The laboratory setting provides a neutral and 'safe' space for the participants to create a shared understanding of complex and endemic issues. The following characteristics of the learning lab provide a powerful catalyst for alignment of divergent mental models in the organization.

- The laboratory environment is neutral and non-threatening. The emphasis is on learning and theory building (what we *don't* know), not on winning or display of knowledge.
- Lack of hierarchy. Managers and staff are equal in this environment. The traditional hierarchy is minimized in the laboratory setting.
- The response time is fast. Hence, the feedback cycle is short, which leads to rapid learning.
- There is no cost or 'loss of face' attached to failure. Hence, it is safe to make mistakes. In fact, mistakes provide opportunities for learning.
- People can see the consequences of their actions first hand. No one attempts to convince or teach anyone else or force his or her preconceived views on others. People learn by themselves and through group interactions.

Implications for Management

The practice field and the learning lab concepts offer fresh and challenging implications for managers and their role. They suggest that a leader/manager should think as a *scientist*, be open to and welcome hard questions, experiment with new ideas, and be prepared to be *wrong*. This requires managers to learn systems thinking skills and use them not just for 'solving' problems but as powerful tools for communication, team building and organizational learning. This means that an effective leader should be the 'designer' of the ship and not its captain [31]. Once they have designed a new structure, strategy, policy or procedure then the managers/leaders should allow (i. e. create a practice field for) the staff to experience the new design, and experiment with it and learn for themselves – the desired outcome is shared understanding leading to alignment of thoughts and actions. This is the essence of organizational learning.

Future Directions

Agent-Based Modeling (ABM)

Agent based modeling (ABM) is an emerging modeling technology which draws its theories and techniques from complexity science [29]. While System Dynamics and Agent-Based Modeling (ABM) use different modeling philosophies and approaches, they can be used complementarily and synergistically.

System Dynamics focuses on modeling structures (i. e. relationships, policies, strategies) that underlie behavior of systems. This may be viewed as a weakness of system dynamics approach in that behavior is assumed to be solely a function of structure (model relationships defined a priori). In contrast, in ABM, organizations are modeled as a system of semi-autonomous decision-making elements – purposeful individuals called *agents*. Each agent individually assesses its situation and makes decisions based upon value hierarchies representing goals, preferences, and standards for behavior. Thus, macro-behavior is not modeled separately but *emerges* from the micro-decisions of individual agents. In other words, in agent based modeling; "emergent" behavior is expected as a result of agents' interactions. This is a key difference between the two approaches.

While system dynamics acknowledges the critical role of individual and organizational mental models (e. g., motivations, values, norms, biases, etc.) it does not explicitly model them. SD utilizes factual data or "cold knowledge" and does not take into account decision makers 'mood'. In contrast, ABM attempts to capture "warm knowledge", representing emotional and human context of decision-making.

Recent advances in video game technology allow the development of multi-agent, artificial 'society' simulators with capabilities for modeling physiology, stress and emotion in decision-making [34]. At the simplest level, an agent-based model consists of a system of agents and their relationships. This new approach enables superior understanding of the complexity in organizations and their relevant business environments. This in turn provides an opportunity for new sophistications in game-play that enhances decision-making. Experience with agent-based modeling shows that even a simple agent-based model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real world system that emulates them.

Despite their differences, SD and ABM can be used in a complementary fashion. Both ABM and SD are powerful tools for transforming information into knowledge and understanding leading to individual and group learning. However, the transition from knowledge to understanding may not be immediate or transparent. This requires a deep shift in mental models through experimentation and group learning.

Systems Thinking and Sustainability

Systems Thinking has a natural affinity with sustainability modeling and management. Sustainability issues are complex; cut across several disciplines; involve multiple stakeholders and require a long term integrated approach. Thus, the systems paradigm and tools have direct and powerful applications in sustainability issues and management.

The applications of system dynamics in sustainability go back to the early 1970s with Jay Forrester's "World2 and World3 analyzes against 30 years of history", followed by "World Dynamics" and "Limits to Growth" [24] and "Beyond the Limits" [25]. "The politics of the environment has also evolved dramatically since 1970. Public awareness of the reality of the environmental challenge has risen; Ministries of Environment have become commonplace" ([28] p. 220). As an example today concern over carbon emissions has already become an international currency. As a result, sustainability has brought a fresh challenge for governments, business and industry, scientists, farmers and all the citizens of the world collectively to find systemic solutions that are mutually and globally agreeable. Systems Thinking and System dynamics can make real and valuable contributions to addressing this challenge.

Bibliography

Primary Literature

1. Ackoff RA (1999) Re-creating the corporation – A design of organizations for the 21st century. Oxford University Press, Oxford
2. Argyris C (1992) The next challenge for TQM: Overcoming organisational defences. J Qual Particip 15:26–29
3. Argyris C, Schon D (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Bontis N, Crossan M, Hulland J (2002) Managing an organizational learning system by aligning stocks and flows. J Manag Stud 39(4):437–469
5. Brown JS (1991) Research that reinvents the corporation. Harv Bus Rev 68:102–111
6. Cavana R, Maani K (2004) A methodological framework for integrating systems thinking and system dynamics. In: System Dynamics Society Proceedings. Oxford
7. Cohen E, Tichy N (1997) How leaders develop leaders. Training and Development, May
8. Common R (2004) Organisational learning in a political environment: Improving policy-making in UK government. Policy Stud 25(1):35–49
9. De Geus A (1997) The living company. Harv Bus Rev 75(2):51–59
10. Flood RL (1999) Rethinking the fifth discipline: Learning within the unknowable. Routledge, London
11. Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge
12. Forrester JW (1971) World dynamics. Wright-Allen (Subsequently re-published by Productivity Press, and Pegasus Communications)
13. Forrester JW (1994) Building a foundation for tomorrow's organizations. In: Systems thinking in action video collection, vol 1. Pegasus Communications, Cambridge
14. Goodstein L, Burke W (1991) Creating successful organisation change. Organ Dyn 19(4):5–17
15. Imants J (2003) Two basic mechanisms for organizational learning in schools. Europ J Teach Educ 26(3):293–311
16. Isaacs W (1993) Taking flight: Dialogue, collective thinking and organisational learning. Organ Dyn 22(2):24–39
17. Kim DH (1993) The link between individual and organizational learning. Sloan Manag Rev 35(1):37–50
18. Maani K (2002) Consensus building through systems thinking – the case of policy and planning in healthcare. Aust J Inform Syst 9(2):84–93
19. Maani K, Benton C (1999) Rapid team learning. Lessons from team New Zealand's America's cup campaign. Organ Dyn 27(4)
20. Maani K, Cavana R (2007) Systems methodology. Syst Think 18(8):2–7
21. Maani K, Cavana R (2007) Systems thinking, system dynamics – Managing change and complexity, 2nd edn. Prentice Hall, Pearson Education, Auckland
22. Maani K, Pourdehnad J, Sedehi H (2003) Integrating system dynamics and intelligent agent-based modelling – theory and case study. Euro INFORMS, Istanbul
23. March JG, Olsen JP (1975) The uncertainty of the past; Organizational ambiguous learning. Europ J Polit Res 3:147–171
24. Meadows DH, Meadows DL, Randers J, Behren W (1972) The limits to growth. Universe Press, New York
25. Meadows DH, Meadows DL, Randers J (1992) Beyond the limits. Chelsey Green, Post Mills
26. Nadler DA, Tushman ML (1990) Beyond the charismatic leader: Leadership and organisational change. Calif Manag Rev
27. Nonaka I, Takeuchi H (1995) The knowledge creating company. Oxford University Press, New York
28. Randers J (2000) From limits to growth to sustainable development or SD (sustainable development) in a SD (system dynamics) perspective. Syst Dyn Rev 16(3):213–224
29. Rothfeder J (2003) Expert voices: Icosystem's Eric Bonabeau. CIO Insights
30. Senge P (1990) The fifth discipline: The art and practice of the learning organisation. Currency
31. Senge P (1990) The leader's New Work: Building learning organisation's. Sloan Manag Rev:7–23
32. Senge P (1992) Building learning organisation's. J Qual Particip:1–8
33. Senge P, Sterman JD (1991) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. In: Kochan T, Useem M (eds) Transforming organizations. Oxford University Press, Oxford
34. Silverman BG et al (2002) Using human models to improve the realism of synthetic agents. Cogn Sci Q 3
35. Sterman JD (2000) Business dynamics, systems thinking and modeling for a complex world. McGraw-Hill, Irwin
36. System Dynamics Society website http://www.systemdynamics.org/
37. US Department of Energy Introduction to system dynamics, A systems approach to understanding complex policy issues, US Department of Energy. http://www.systemdynamics.org/DL-IntroSysDyn/inside.htm
38. Vennix JAM (1995) Building consensus in strategic decision-making: System Dynamics As A Support System. Group Decis Negot 4(4):335–355
39. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester, chapt 5

40. Wolstenholme E (1997) System dynamics in the elevator (SD1163), e-mail communication, 24 Oct 1997 systemdynamics@world.std.com

Books and Reviews

(This section is due to M. Anjali Sastry and John D. Sterman, "An Annotated Survey of the Essential System Dynamics Literature System Dynamics Group", Sloan School of Management, MIT)

Industrial and Economic Dynamics: The Foundations

Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge (Presents dynamic analysis of a business problem through a model of a production-distribution system that shows oscillatory behavior. Policies to improve system performance are discussed, and numerous policy experiments are demonstrated. Includes full equation listing.)

Forrester JW (1968) Principles of systems. Productivity Press, Cambridge (System structure and behavior are differentiated, with examples showing how structure determines behavior. Rates and levels are described. Inventory model shows effects of delivery delay and resulting production cycles.)

Forrester JW (1975) Collected papers of Jay W. Forrester. Productivity Press, Cambridge (Includes many seminal papers, such as Industrial Dynamics: A Major Breakthrough for Decision Makers; Common Foundations Underlying Engineering and Management; A New Corporate Design; Market Growth as Influenced by Capital Investment; and Counterintuitive Behavior of Social Systems.)

Forrester JW (1989) The beginnings of system dynamics (Working Paper No. D-4165). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A personal history beginning on the high plains of western Nebraska. Describes the early projects that shaped the field.)

Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge (Shows how production scheduling and work force management policies generate the 3–5 year business cycle. Economic cycles, in turn, are caused by capital investment policies that fail to account for delays in acquiring long-lead time plant and equipment.)

Meadows DL (1970) Dynamics of commodity production cycles. Productivity Press, Cambridge (Develops a simple generic model of commodity supply and demand with explicit production capacity and delays, prices and markets. Applies the model to hogs, chicken and cattle.)

Urban and Public Policy Dynamics

Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Productivity Press, Cambridge (A very readable introductory text. Uses the urban system as an example to teach general points about modeling methods, formulation and analysis.)

Forrester JW (1969) Urban dynamics. Productivity Press, Cambridge (Seminal model of urban growth and decay, controversial then and vindicated now. Chapter 6 describes general characteristics of complex systems such as compensating feedback and shifting the burden to the intervener.)

Mass NJ (ed) (1974) Readings in urban dynamics, vol I. Productivity Press, Cambridge (Extensions, modification, and responses to criticisms of the Urban Dynamics model.)

Schroeder WW, III, Sweeney RE, Alfeld LE (eds) (1975) Readings in urban dynamics, vol II. Productivity Press, Cambridge (Further extends and explores the Urban Dynamics model.)

Limits to Growth and Other Global Models

Forrester JW (1973) World Dynamics, 2nd edn. Productivity Press, Cambridge (The first global model, on which Limits to Growth was based. The extreme simplicity of the model allowed it to be presented to a wide audience.)

Meadows DL, Meadows DH (eds) (1974) Toward global equilibrium: Collected papers. Productivity Press, Cambridge (Describes and explores, through system dynamics models, policies for sustainability designed to avoid the collapse shown in the 'business as usual' WORLD3 scenarios.)

Meadows DH, Meadows DL, Randers J, Behrens WW III (1972) The limits to growth: A report for the club of Rome's project on the predicament of mankind. Universe Books, New York (Classic controversial study of the human future. Nontechnical presentation of structure, assumptions, and results of the WORLD3 model. Concluded that present policies were unsustainable; shows how alternate policies could stabilize population at a high standard of living.)

Meadows DL, Behrens WW III, Meadows DH, Naill RF, Randers J, Zahn EKO (1974) Dynamics of growth in a finite world. Productivity Press, Cambridge (Full documentation and data for the WORLD3 model used in the Limits to Growth. Describes the structure and assumptions; includes all data needed for complete replication of all runs in the popular book. Formulations described here may be useful to all system dynamics modelers.)

Meadows D, Richardson J, Bruckmann G (1982) Groping in the dark. Wiley, New York (Describes a range of global models built under different approaches and discusses the strengths, weaknesses, and implications of each. Presented in an engaging, personal style.)

Meadows DH, Meadows DL Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green, Post Mills (Follows up on Limits to Growth. Shows that many problems described in 1972 have worsened, as predicted by the model. Argues for a shift in values necessary to create a sustainable and equitable future.)

SD for Management: Firm and Market Models

Coyle RG (1977) Management System Dynamics. Wiley, New York (Text emphasizing managerial modeling, with a focus on operations and examples including discrete elements.)

Hall RI (1976) A system pathology of an organization: The rise and fall of the Old Saturday Evening Post. Adm Sci Q 21(2):185–211 (A case-study using a system dynamics model to explain how failure to understand the feedbacks among policies governing ad rates, ad and editorial pages, marketing, and pricing lead to the failure of the Post just as circulation reached an all-time high.)

Lyneis JM (1980) Corporate planning and policy design. Productivity Press, Cambridge (Begins with a simple model of inventory management in a manufacturing firm and gradually extends the model to one of the entire firm.)

Merten PP (1991) Loop-based strategic decision support systems. Strat Manag J 12:371–382 (Describes a model of a multinational firm establishing new markets in less-developed countries. Captures qualitative shifts in firm structure and organization endogenously as the firm evolves.)

Morecroft JDW (1984) Strategy support models. Strat Manag J 5(3):215–229 (Describes the use of models as participants in the ongoing dialogue among managers regarding strategy formation and evaluation. Emphasizes the processes for model development and use that enhance the utility of modeling in design of high-level corporate strategy.)

Morecroft JDW, Lane DC, Viita PS (1991) Modelling growth strategy in a biotechnology startup firm. Syst Dyn Rev 7(2):93–116 (Describes a case-study of a start-up in which system dynamics modeling helps to define a desirable growth strategy for the firm. The integrated model generated strategies that allowed different parts of the firm to choose consistent approaches.)

Roberts EB (ed) (1978) Managerial applications of system dynamics. Productivity Press, Cambridge (Extensive collection of early corporate models, including history and commentary by practitioners. Covers R&D management, production and operations, human resources, and other applications areas.)

Economic Models

Forrester JW (1989) The system dynamics national model: Macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin (Provides an overview of the national modeling project in which both micro- and macro-economic factors are included. Model generates business cycles, inflation, stagflation, the economic long wave, and growth.)

Saeed K (1986) The dynamics of economic growth and political instability in the developing countries. Syst Dyn Rev 2(1):20–35 (Shows how rapid economic development can generate social and political instability through a model that links socio-political factors to economic development.)

Sterman JD (1985) A behavioral model of the economic long wave. J Econ Behav Organ 6(1):17–53 (Proposes and tests a simple model of the long wave. The intended rationality of each decision rule is tested and the long wave is explained as the unintended result of the interaction of locally rational decision processes. The model is the basis for the STRATAGEM-2 game, and can exhibit chaos.)

Sterman JD (1989) Deterministic chaos in an experimental economic system. J Econ Behav Organ 12:1–28 (Sterman's 1985 model of the long wave is converted into a management flight simulator and used as an experiment in which subjects make the capital investment decision. Simple decision rules capturing subject's policies are estimated and explain their behavior well. Simulation of these rules yields deterministic chaos for about 25% of the subjects.)

Sterman JD (1986) The economic long wave: Theory and evidence. Syst Dyn Rev 2(2):87–125 (Comprehensive overview of the theory of long waves arising from the System Dynamics National Model. Reviews the feedback structures responsible for the long wave and empirical evidence supporting the dynamic hypotheses. Discusses the role of innovation and political value change.)

Conceptualizing, Formulating and Validating Models

Barlas Y (1989) Multiple tests for validation of system dynamics type of simulation models. Europ J Operat Res 42(1):59–87 (Discusses a variety of tests to validate SD models, including structural and statistical tests.)

Barlas Y, Carpenter S (1990) Philosophical roots of model validation: Two paradigms. Syst Dyn Rev 6(2):148–166 (Contrasts the system dynamics approach to validity with the traditional, logical empiricist view of science. Finds that the relativist philosophy is consistent with SD and discusses the practical implications for modelers and their critics.)

Forrester JW (1980) Information sources for modeling the national economy. J Am Stat Assoc 75(371):555–574 (Argues that modeling the dynamics of firms, industries, or the economy requires use of multiple data sources, not just numerical data and statistical techniques. Stresses the role of the mental and descriptive data base; emphasizes the need for first-hand field study of decision making.)

Forrester JW (1985) The model versus a modeling process. Syst Dyn Rev 1(1):133–134 (The value of a model lies not in its predictive ability alone but primarily in the learning generated during the modeling process.)

Forrester JW (1987) Fourteen 'Obvious Truths'. Syst Dyn Rev 3(2):156–159 (The core of the system dynamics paradigm, as seen by the founder of the field.)

Forrester JW (1987) Nonlinearity in high-order models of social systems. Europ J Operat Res 30(2):104–109 (Nonlinearity is pervasive, unavoidable, and essential to the functioning of natural and human systems. Modeling methods must embrace nonlinearity to yield realistic and useful models. Linear and nearly-linear methods are likely to obscure understanding or lead to erroneous conclusions.)

Homer JB (1983) Partial-model testing as a validation tool for system dynamics. In: International System Dynamics Conference, pp 920–932 (How model validity can be improved through partial model testing when data for the full model are lacking.)

Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System dynamics. In: TIMS studies in the management sciences, vol 14. North-Holland, Amsterdam (Collection of papers focused on methodology. Includes Forrester and Senge on Tests for Building Confidence in System Dynamics Models and Gardiner & Ford's discussion on Which Policy Run is Best, and Who Says So?)

Mass N (1991) Diagnosing surprise model behavior: A tool for evolving behavioral and policy insights. Syst Dyn Rev 7(1):68–86 (Guidelines for learning from surprise model behavior with tests to resolve anomalous behavior.)

Morecroft JDW (1982) A critical review of diagramming tools for conceptualizing feedback system models. Dynamica 8(1):20–29 (Critiques causal-loop diagrams and proposes subsystem and policy structure diagrams as superior tools for representing the structure of decisions in feedback models.)

Randers J (ed) (1980) Elements of the system dynamics method. Productivity Press, Cambridge (Includes Mass on Stock and Flow Variables and the Dynamics of Supply and Demand; Mass & Senge on Alternative Tests for Selecting Model Variables; and Randers' very useful Guidelines for Model Conceptualization.)

Richardson GP (1986) Problems with causal-loop diagrams. Syst Dyn Rev 2(2):158–170 (Causal-loop diagrams cannot show stock-and-flow structure explicitly and can obscure important dynamics. Offers guidelines for proper use and interpretation of CLDs.)

Richardson GP, Pugh AL III (1981) Introduction to system dynamics modeling with DYNAMO. Productivity Press, Cambridge (Introductory text with excellent treatment of conceptualization, stocks and flows, formulation, and analysis. A good way to learn the DYNAMO simulation language as well.)

Roberts N, Andersen DF, Deal RM, Grant MS, Shaffer WA (1983) Introduction to computer simulation: A system dynamics modeling approach. Addison-Wesley, Reading (Easy-to-understand introductory text, complete with exercises.)

Sterman JD (1984) Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. Dynamica 10(2):51–66 (Describes the use of rigorous statistical tools for establishing model validity. Shows how Theil statistics can be used to assess goodness-of-fit in dynamic models.)

Wolstenholme EF (1990) System enquiry – A system dynamics approach. Wiley, Chichester (Describes a research methodology for building a system dynamics analysis. Emphasizes causal-loop diagramming, mapping of mental models, and other tools for qualitative system dynamics.)

Modeling for Learning: Systems Thinking and Organizational Learning

Kim D (1989) Learning laboratories: Designing a reflective learning environment. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International system dynamics conference. Springer, Berlin (A case-study of a process designed to convey dynamic insights to participants in a workshop setting designed around a management flight simulator game.)

Morecroft JDW (1988) System dynamics and microworlds for policymakers. Europ J Operat Res 35(3):301–320 (Describes the model-building tools available to managers and policymakers.)

Morecroft JDW, Sterman JD (eds) (1992) Modelling for Learning. Eur J Operat Res Special Issue 59(1) (17 papers describing models and methods to enhance learning, both for individuals and organizations. Covers elicitation and group process techniques, management flight simulators, and tools for capturing, representing, and simulating mental and formal models.)

Richmond B (1990) Systems thinking: A critical set of critical thinking skills for the 90's and beyond. In: Andersen DF, Richardson GP, Sterman JD (eds) International System Dynamics Conference, 1990 (Proposes a process and skill set to teach systems thinking. The process relies on learner-directed learning. The skill set includes general scientific reasoning and SD, supported by simulation.)

Senge PM (1990) Catalyzing systems thinking within organizations. In: Masarik F (ed) Advances in organization development. Ablex, Norwood (Presents a case study in which the use of system dynamics generated insights into a chronic business problem. Steps in generating, testing and disseminating a system dynamics model are described.)

Senge PM (1990) The fifth discipline: The art and practice of the learning organization. Doubleday Currency, New York (Introduces systems thinking as part of a wider approach to organizational learning. Conveys basic system structures to a non-technical business audience by means of anecdotes and archetypes.)

Decision Making

Morecroft JDW (1983) System dynamics: Portraying bounded rationality. Omega 11(2):131–142 (SD models represent decision making as boundedly rational. Reviews and contrasts the concept of bounded rationality as developed by Herbert Simon. Uses Forrester's Market Growth model to show how locally rational decision rules can interact to yield globally dysfunctional outcomes.)

Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916 (Shows how the intended rationality of decision rules in SD models can be assessed, and how one analyzes a simulation model and output to understand the assumed bounds on rationality in dynamic models. A model of salesforce effort allocation is used to illustrate.)

Sterman JD (1987) Expectation formation in behavioral simulation models. Behav Sci 32:190–211 (Proposes and tests a simple dynamic model of expectation formation in dynamic models (the TREND function). Shows how the TREND function explains a forty year history of inflation forecasts and several different types of long-term energy demand forecasts.)

Sterman JD (1989) Misperceptions of feedback in dynamic decision making. Organ Behav Hum Decis Process 43(3):301–335 (Describes an experiment with a simple economic system in which subjects systematically generate costly oscillations. Estimates decision rules to characterize subject behavior. Finds that people systematically ignore feedbacks, time delays, accumulations, and nonlinearities. These misperceptions of feedback lead to poor quality decisions when dynamic complexity is high.)

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339 (Analyzes the results of the Beer Distribution Game. Misperceptions of feedback are found to cause poor performance in the beer game, as in other experiments. Estimates of the subjects' decision rules show they ignore time delays, accumulations, feedbacks, and nonlinearities.)

Selected Applications of SD

Abdel-Hamid TK, Madnick SE (1991) Software project dynamics: An integrated approach. Prentice Hall, Englewood Cliffs (Integrated SD model of the software development process. The model covers design, coding, reviewing, and quality assurance; these are integrated with resource planning, scheduling, and management of software projects. Includes full documentation, validation, and policy tests.)

Cooper KG (1980) Naval ship production: A claim settled and a framework built. Interfaces 10(6) (An SD model was used to quantify the causes of cost overruns in a large military shipbuilding project. One of the first and most successful applications of system dynamics to large-scale project management; initiated a long line of related project modeling work.)

Ford A, Bull M (1989) Using system dynamics for conservation policy analysis in the pacific northwest. Syst Dyn Rev 5(1):1–15 (Describes the use of an extensive SD model of electric power generation with endogenous demand. The model is used to evaluate strategies for conservation and new generation capacity. Includes discussion of implementation and integration of the SD model with other existing planning tools.)

Gardiner LK, Shreckengost RC (1987) A system dynamics model for estimating heroin imports into the United States. Syst Dyn Rev 3(1):8–27 (Describes how the CIA used SD to estimate the illegal importation of drugs to the US.)

Homer JB (1985) Worker burnout: A dynamic model with implications for prevention and control. Syst Dyn Rev 1(1):42–62 (Explains how knowledge workers can experience cycles of burnout through a simple system dynamics model. Avoiding burnout requires that one work at less than maximum capacity.)

Homer JB (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Chang 31(3):197–218 (Presents a generic model of the diffusion of new medical technologies. Case studies of the cardiac pacemaker and an antibiotic illustrate how the same model can explain the different diffusion dynamics of successful and unsuccessful technologies.)

Homer JB (1993) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78 (An excellent model of the interacting dynamics of addiction, policy-setting, and enforcement.)

Jensen KS, Mosekilde E, Holstein-Rathlou N (1985) Self-sustained oscillations and chaotic behaviour in kidney pressure regulation. In: Prigogine I, Sanglier M (eds) Laws of nature and human conduct. Taskforce of Research Information and Study on Science, Brussels (Presents a system dynamics model of the dynamics of rat kidneys. Experimental data show previously unexplained oscillations, sometimes chaotic. The model explains how these fluctuations arise. Excellent example of SD applied to physiology.)

Levin G, Hirsch GB, Roberts EB (1975) The persistent poppy: A computer-aided search for heroin policy. Ballinger, Cambridge (Examines the interactions within a community among drug users, the police and justice system, treatment agencies, and the citizens. Analyzes policies designed to restore the community's health.)

Levin G, Roberts EB, Hirsch GB, Kligler DS, Roberts N, Wilder JF (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge (Presents a generic theory of human service delivery, with case studies and examples drawn from mental health care, dental planning, elementary education, and outpatient care.)

Naill RF (1992) A system dynamics model for national energy policy planning. Syst Dyn Rev 8(1):1–19

Naill RF, Belanger S, Klinger A, Peterson E (1992) An analysis of the cost effectiveness of US energy policies to mitigate global warming. Syst Dyn Rev 8(2):111–128 (Reviews the 20 year history of the SD energy models used by the US Dept. of Energy to forecast and analyze policy options for national energy security, including the impact of US policies on global climate change.)

Sklar Reichelt K (1990) Halter marine: A case study of the dangers of litigation. (Working Paper No. D-4179). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A case-study illustrating the use of system dynamics in litigation. Suitable for classroom teaching.)

Sturis J, Polonsky KS, Mosekilde E, Van Cauter E (1991) Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. Am J Physiol 260(Endocrinol. Metab. 23):E801–E809 (New experimental data show that the human glucose/insulin system is inherently oscillatory. An SD model explains these dynamics. The model is validated against detailed physiological data.)

Cross-Fertilization and Comparative Methodology

Allen PM (1988) Dynamic models of evolving systems. Syst Dyn Rev 4(1–2):109–130 (Reviews approaches to nonlinear dynamics, self-organization, and evolution developed in the Brussels school by Prigogine, Allen, and others. Provides illustrations and examples.)

Kim DH (1990) Toward learning organizations: Integrating total quality control and systems thinking. (Working Paper No. D-4036). System Dynamics Group, Sloan School of Management, MIT, Cambridge (Argues that SD and Total Quality Management are complementary approaches to improvement and organizational learning. Systems thinking and modeling are needed to speed the improvement cycle for processes with long time delays.)

Meadows DH, Robinson JM (1985) The electronic oracle: Computer models and social decisions. Wiley (Comparative assessment of the underlying assumptions, boundary, limitations, and uses of different models, including optimization, simulation, and econometrics. Offers guidelines for assessing model assumptions, including ways to recognize the implicit biases of each modeling paradigm.)

Powers WT (1990) Control theory: A model of organisms. Syst Dyn Rev 6(1):1–20 (An explicit feedback control perspective on perception and decision making in living organisms. Argues the behaviorist and cognitive paradigms have fundamentally misunderstood the concept of feedback. For Powers, feedback allows organisms to control perceptions by altering behavior.)

Radzicki MJ (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147 (Surveys the institutionalist paradigm in economics and argues that system dynamics is compatible with the institutionalist perspective. The SD approach offers a means by which institutional theories can be formalized and tested.)

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Chang 28(2):93–122 (Presents a formal dynamic model of TS Kuhn's theory of scientific revolutions.)

Sterman JD (1988) A skeptic's guide to computer models. In: Grant L, Lanham MD (eds) Foresight and national decisions. University Press of America (Reviews different modeling methods and their underlying assumptions in nontechnical language. Provides a list of questions model users should ask to assess whether a model or method are appropriate to the problem.)

Other Themes: Pulling the Threads Together

Cooper K, Steinhurst W (eds) (1992) The system dynamics society bibliography. System Dynamics Society. Available from Julie Pugh, 49 Bedford Rd., Lincoln MA, USA 01773. (Lists over 3,000 system dynamics journal articles, books, conference proceedings and working papers. Available in computer-readable format and compatible with bibliographic software)

Meadows DH (1989) System dynamics meets the press. Syst Dyn Rev 5(1):68–80 (Reviews the history of encounters between SD and the media. Offers guidelines for effective communication to the public at large. Stresses the importance of communicating even the simplest system concepts.)

Meadows DH (1991) The global citizen. Island Press, Washington (A collection of Dana's syndicated newspaper columns applying system dynamics principles to problems of everyday life, from organic farming to the fall of the Soviet Union. Emphasizes environmental issues.)

Richardson GP (1991) Feedback thought in social science. University of Pennsylvania Press (Traces the history of the concept of feedback in the social sciences through two threads of thought – the cybernetic and feedback threads. System dynamics is placed in context in a readable and scholarly manner.)

Software

DYNAMO. Pugh-Roberts Associates, Cambridge MA. (The first widely-used computer language developed to simulate system dynamics models, DYNAMO is still in use, available for mainframes and PCs. Many of the models in the system dynamics literature were simulated in DYNAMO)

DYSMAP. University of Salford, UK (PC-based simulation language with syntax similar to DYNAMO. Includes optimization capability based on hill-climbing.)

Microworld Creator and S^4. Microworlds Inc., Cambridge MA (Easy to use environment for simulation and gaming. S^4, the 'industrial strength' version, supports arrays and includes diagnostics for analyzing behavior. Both Creator and S^4 support user-defined information displays and facilitate rapid development of management flight simulators.)

STELLA and ithink. High Performance Systems, Hanover NH. (User-friendly modeling software with full graphical interface. Models are entered graphically, at the level of the stock and flow diagram. Widely used in education from elementary school up; also used in research and practice.)

Vensim. Ventana Systems, Harvard MA. (Powerful simulation environment for SD models. Runs on workstations and PCs. Includes array capability and a wide range of features for analyzing model behavior.)

System Dynamics in the Evolution of the Systems Approach

MARKUS SCHWANINGER
Institute of Management, University of St. Gallen,
St. Gallen, Switzerland

Article Outline

Glossary

Cybernetics The science of communication and control in complex, dynamical systems. The core objects of study are information, communication, feedback and adaptation. In the newer versions of cybernetics, the emphasis is on observation, self-organization, self-reference and learning.

Dynamical system The dynamical system concept is a mathematical formalization of time-dependent processes. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a river, and the evolution of a population of fish in a lake.

Law of requisite variety Ashby's law of requisite variety says: "Only variety can destroy variety". It implies that the varieties of two interacting systems must be in balance, if stability is to be achieved.

Organizational cybernetics The science which applies cybernetic principles to organization. Synonyms are *Management Cybernetics* and *Managerial Cybernetics*.

System There are many definitions of *system*. Two examples: A portion of the world sufficiently well defined to be the subject of study; something characterized by a structure, for example, a social system (Anatol Rapoport). A system is a family of relationships between its members acting as a whole (International Society for the Systems Sciences).

System dynamics A methodology and discipline for the modeling, simulation and control of dynamic systems. The main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, which are modeled as networks of informationally closed feedback loops between stock and flow variables.

Systems approach A perspective of inquiry, education and management, which is based on system theory and cybernetics.

System theory A formal science of the structure, behavior, and development of systems. In fact there are different system theories. General system theory is a transdisciplinary framework for the description and analysis of any kind of system. System theories have been developed in many domains, e. g., mathematics, computer science, engineering, sociology, psychotherapy, biology and ecology.

Variety A technical term for *complexity* which denotes the number of (potential) states of a system.

Definition of the Subject

The purpose of this chapter is to give an overview of the role of system dynamics (SD) in the context of the evolution of the systems movement. This is necessary because SD is often erroneously taken as the systems approach as such, not as part of it. It is also requisite to show that the processes of the evolution of both SD in particular and the systems movement as a whole are intimately linked and intertwined. Finally, in view of the purpose of the chapter the actual and potential relationships between system dynamics and the other strands of the systems movement are evaluated. This way, complementarities and synergies are identified.

Introduction

The purpose of this contribution is to give an overview of the role of system dynamics in the context of the evolution of the systems movement. "Systems movement" – often referred to briefly as "systemics" – is a broad term, which takes into account the fact that there is no single system approach, but a range of different ones. The common denominator of the different system approaches in our day is that they share a worldview focused on complex dynamic systems, and an interest in describing, explaining and designing or at least influencing them. Therefore, most of the system approaches offer not only a theory but also a way of thinking ("systems thinking" or "systemic thinking") and a methodology for dealing with systemic issues or problems.

System dynamics (SD) is a discipline and a methodology for the modeling, simulation and control of complex, dynamic systems. SD was developed by MIT professor Jay W. Forrester (e. g. [20,21]) and has been propagated by his students and associates. SD has grown to a school of numerous academics and practitioners all over the world. The particular approach of SD lies in representing the issues or systems-in-focus as meshes of closed feedback loops made up of stocks and flows, in continuous time and subject to delays.

The development of the system dynamics methodology and the worldwide community that applies SD to modeling and simulation in radically different contexts suggest that it is a "systems approach" on its own. Nevertheless, taking "system dynamics" as the (one and only) synonym for "systemic thinking" would be going too far, given the other approaches to systemic thinking as well as a variety of system theories and methodologies, many of which are complementary to SD. In any case, however, the SD community has become the strongest "school" of the Systems approach, if one takes the numbers of members in organizations representing the different schools as a measure (by 2006, the System Dynamics Society had more than 1000 members).

The rationale and structure of this contribution is as follows. Starting with the emergence of the systems approach, the multiple roots and theoretical streams of systemics are outlined. Next, the common grounds and differences among different strands of the systems approach are highlighted, and the various systems methodologies are explored. Then the distinctive features of SD are analyzed. Finally comes a reflection on the relationships of SD with the rest of the systems movement as well as with potential complementarities and synergies.

In Table 1, a time-line overview of some milestones in the evolution of the systems approach in general and System Dynamics in particular is given. Elaborating on each of the sources quoted therein would reach beyond the purpose of this chapter. However, to convey a synoptic view, a diagram showing the different systems approaches and their interrelationships is provided in the Appendix "Systems Approaches – An Overview".

Emergence of the Systems Approach

The systems movement has many roots and facets, with some of its concepts going back as far as ancient Greece. What we name as "the systems approach" today materialized in the first half of the twentieth century. At least two important components should be mentioned: those proposed by von Bertalanffy and by Wiener.

Ludwig von Bertalanffy, an American biologist of Austrian origin, developed the idea that organized wholes of any kind should be describable and, to a certain extent, explainable, by means of the same categories, and ultimately by one and the same formal apparatus. His *general systems theory* triggered a whole movement which has tried to identify invariant structures and mechanisms across different kinds of organized wholes (for example, hierarchy, teleology, purposefulness, differentiation, morphogenesis, stability, ultrastability, emergence, and evolution).

In 1948 Norbert Wiener, an American mathematician at the Massachusetts Institute of Technology, published his seminal book on *Cybernetics*, building upon interdisciplinary work carried out in cooperation with Bigelow, an IBM engineer, and Rosenblueth, a physiologist. Wiener's opus became the transdisciplinary foundation for a new science of capturing as well as designing control and communication mechanisms in all kinds of dynamic systems [81]. Cyberneticists have been interested in concepts such as information, communication, complexity, autonomy, interdependence, cooperation and conflict, self-production ("autopoiesis"), self-organization, (self-) control, self-reference and (self-) transformation of complex dynamic systems.

Along the genetic line of the tradition which led to the evolution of General Systems Theory (von Bertalanffy, Boulding, Gerard, Miller, Rapoport) and Cybernetics (Wiener, McCulloch, Ashby, Powers, Pask, Beer), a number of roots can be identified, in particular:

- Mathematics (for example, Newton, Poincaré, Lyapunov, Lotka, Volterra, Rashevsky)
- Logic (for example, Epimenides, Leibniz, Boole, Russell and Whitehead, Goedel, Spencer-Brown)
- Biology, including general physiology and neurophysiology (for example, Hippocrates, Cannon, Rosenblueth, McCulloch, Rosen)
- Engineering and computer science, including the respective physical and mathematical foundations (for example, Heron, Kepler, Watt, Euler, Fourier, Maxwell, Hertz, Turing, Shannon and Weaver, von Neumann, Walsh)
- Social and human sciences, including economics (for example, Hume, Adam Smith, Adam Ferguson, John Stuart Mill, Dewey, Bateson, Merton, Simon, Piaget).

In this last-mentioned strand of the systems movement, one focus of inquiry is on the role of feedback in communication and control in (and between) organizations and society, as well as in technical systems. The other focus of interest is on the multidimensional nature and the multi-level structures of complex systems. Specific theory building, methodological developments and pertinent applications have occurred at the following levels:

- Individual and family levels (for example, systemic psychotherapy, family therapy, holistic medicine, cognitive therapy, reality therapy)
- Organizational and societal levels (for example, managerial cybernetics, organizational cybernetics, sociocybernetics, social systems design, social ecology, learning organizations)
- The level of complex (socio-)technical systems (systems engineering)

The notion of "socio-technical systems" has become widely used in the context of the design of organized wholes involving interactions of people and technology (for instance, Linstone's multi-perspectives-framework, known by way of the mnemonic TOP (*T*echnical, *O*rganizational, *P*ersonal/individual).

As can be noted from these preliminaries, different kinds of system theory and methodology have evolved over time. One of these is a theory of dynamic systems by Jay W. Forrester, which serves as a basis for the methodology of system dynamics. Two eminent titles are [20] and [21]. In SD, the main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, modeled as networks of informationally closed feedback loops between stock and flow variables. Several other mathematical systems theories have been elaborated, for example, mathematical general systems theory (Klir, Pestel, Mesarovic and Takahara), as well as a whole stream of theoretical developments which can be subsumed under the terms "dynamic systems theory" or "theories of non-linear dynamics" (for example, catastrophe theory, chaos theory and complexity theory). Under the latter, branches such as the theory of fractals (Mandelbrot), geometry of behavior (Abraham), self-organized criticality (Bak), and network theory (Barabasi, Watts) are subsumed. In this context, the term "sciences of complexity" is used.

In addition, a number of mathematical theories, which can be called "system theories," have emerged in different application contexts, examples of which are discernible in the following fields:

- Engineering, namely information and communication theory (Shannon and Weaver), technology and computer-aided systems theory (for example, control theory, automata, cellular automata, agent-based modeling, artificial intelligence, cybernetic machines, neural nets)
- Operations research (for example, modeling theory and simulation methodologies, Markov chains, genetic algorithms, fuzzy control, orthogonal sets, rough sets)
- Social sciences, economics in particular (for example, game theory, decision theory)
- Biology (for example, Sabelli's Bios theory of creation)
- Ecology (for example, E. and H. Odum's systems ecology).

Most of these theories are transdisciplinary in nature, i. e., they can be applied across disciplines. The Bios theory, for example is applicable to clinical, social, ecological and personal settings [54]. Examples of essentially non-mathematical system theories can be found in many different areas of study, e. g.:

- Economics, namely its institutional/evolutionist strand (Veblen, Myrdal, Boulding, Dopfer)
- Sociology (for example, Parsons' and Luhmann's social system theories, Hall's cultural systems theory)
- Political sciences (for example, Easton, Deutsch, Wallerstein)
- Anthropology (for example, Levi Strauss's structuralist-functionalist anthropology, Margaret Mead)
- Semiotics (for example, general semantics (Korzybski, Hayakawa, Rapoport), cybersemiotics (Brier))
- Psychology and psychotherapy (for example, systemic intervention (Bateson, Watzlawick, F. Simon), and fractal affect logic (Ciompi))
- Ethics and epistemology (for example, Vickers, Churchman, von Foerster, van Gigch)

Several system-theoretic contributions have merged the quantitative and the qualitative in new ways. This is the case for example in Rapoport's works in game theory as well as general systems theory, Pask's conversation theory, von Foerster's cybernetics of cybernetics (second-order cybernetics), and Stafford Beer's opus in managerial cybernetics. In all four cases, mathematical expression is virtuously connected to ethical, philosophical, and epistemological reflection. Further examples are Prigogine's theory of dissipative structures, Mandelbrot's theory of fractals, complex adaptive systems (Holland et al.), Kauffman's complexity theory, and Haken's synergetics, all of which combine mathematical analysis and a strong component of qualitative interpretation.

A large number of systems methodologies, with the pertinent threads of systems practice, have emanated from these theoretical developments. Many of them are expounded in detail in specialized encyclopedias (e. g., [27] and, under a specific theme, named *Systems Science and Cybernetics*, of the Encyclopedia of Life Support Systems [18]). In this chapter, only some of these will be addressed explicitly, in order to shed light on the role of SD as part of the systems movement.

Common Grounds and Differences

Even though the spectrum of system theories and methodologies outlined in the preceding section may seem multifarious, all of them have a strong common denominator: They build on the idea of systems as organized wholes. An objectivist working definition of a system is that of a whole, the organization of which is made up by interrelationships. A subjectivist definition is that of a set of interdependent variables in the mind of an observer, or, a mental construct of a whole, an aspect that has been emphasized by the position of constructivism. *Constructivism* is a synonym for *second-order cybernetics*. While first-order cybernetics concentrates on regulation, information and feedback, second-order cybernetics focuses on observation, self-organization and self-reference. Heinz von Foerster established the distinction between 'observed systems' for the former and 'observing systems' for the latter [74].

From the standpoint of operational philosophy, a system is, as Rapoport says, "a part of the world, which is sufficiently well defined to be the object of an inquiry or also something, which is characterized by a structure, for example, a production system" [50].

In recent systems theory, the aspect of relationships has been emphasized as the main building block of a system, as one can see from a definition published by the International Society for the Systems Sciences (ISSS): "A system is a family of relationships between its members acting as a whole" [63]. Also, purpose and interaction have played an important part in reflections on systems: Systems are conceived, in the words of Forrester [21], as "wholes of elements, which cooperate towards a common goal." Purposeful behavior is driven by internal goals, while purposive behavior rests on a function assigned from the outside. Finally, the aspects of open and closed functioning have been emphasized. Open systems are characterized by the import and export of matter, energy and information. A variant of particular relevance in the case of social systems is the operationally closed system, that is, a system which is self-referential in the sense that its self-production (autopoiesis) is a function of production rules and processes by which order and identity are maintained, and which cannot be modified directly from outside. As we shall see, this concept of operational closure is very much in line with the concept of circularity used in SD.

At this point, it is worth elaborating on the specific differences between two major threads of the systems movement, which are of special interest because they are grounded in "feedback thought" [52]: The cybernetic thread, from which organizational cybernetics has emanated, and the servomechanic thread in which SD is grounded. As Richardson's detailed study shows, the strongest influence on cybernetics came from biologists and physiologists, while the thinking of economists and engineers essentially shaped the servomechanic thread. Consequently, the concepts of the former are more focused on the adaptation and control of complex systems for the purpose of maintaining stability under exogenous disturbances. Servomechanics, on the other hand, and SD in particular, take an endogenous view, being mainly interested in understanding circular causality as the principal source of a system's behavior. Cybernetics is more connected with communication theory, the general concern of which can be summarized as how to deal with randomly varying input. SD, on the other hand, shows a stronger link with engineering control theory, which is primarily concerned with behavior generated by the control system itself, and by the role of nonlinearities. Managerial cybernetics and SD both share the concern of contributing to management science, but with different emphases and with instruments that are different but in principle complementary. Finally, the mathematical foundations are generally more evident in the basic literature on SD than in the writings on organizational cybernetics, in which the formal apparatus underlying model formulation is confined to a small number of publications (e. g., [7,10]), which are less known than the qualitative treatises. The terms *management cybernetics* and *managerial cybernetics* are used as synonyms for *organizational cybernetics*.

The Variety of Systems Methodologies

The methodologies that have evolved as part of the systems movement cannot be expounded in detail here. The two epistemological strands in which they are grounded, however, can be identified – the positivist tradition and the interpretivist tradition.

Positivist tradition denotes those methodological approaches that focus on the generation of "positive knowledge," that is, a knowledge based on "positively" ascertained facts. *Interpretivist tradition* denotes those methodological approaches that emphasize the importance of subjective interpretations of phenomena. This stream goes back to Greek art and science of the interpretation and understanding of texts.

Some systems methodologies have been rooted in the positivist tradition, and others in the interpretivist tradition. The differences between the two can be described along the following set of polarities:

- An objectivist versus a subjectivist position
- A conceptual–instrumental versus a communicational/cultural/political rationality
- An inclination to quantitative versus qualitative modeling
- A structuralist versus a discursive orientation.

A positivistic methodological position tends toward the objectivistic, conceptual–instrumental, quantitative and structuralist–functionalist in its approach. An interpretive position, on the other hand, tends to emphasize the subjectivist, communicational, cultural, political, ethical and esthetic—that is, the qualitative and discursive aspects. It would be too simplistic to classify a specific methodology in itself as being "positivistic" or "interpretative". Despite the traditions they have grown out of, several methodologies have evolved and been reinterpreted or opened to new aspects (see below).

In the following, a sample of systems methodologies will be characterized and positioned in relation to these two traditions, beginning with those in the positivistic strand:

- *"Hard" OR methods*. Operations research (OR) uses a wide variety of mathematical and statistical methods and techniques—for example of optimization, queuing, dynamic programming, graph theory, time series analysis—to provide solutions for organizational and manaperial problems, mainly in the operational domains of production and logistics, and in finance.
- *Living systems theory*. In his LST, James Grier Miller [44] identifies a set of 20 necessary components that can be discerned in living systems of any kind. These structural features are specified on the basis of a huge empirical study and proposed as the "critical subsystems" that "make up a living system." LST has been used as a device for diagnosis and design in the domains of engineering and the social sciences.
- *Viable system model*. To date, Stafford Beer's VSM is probably the most important product of organizational cybernetics. It specifies a set of management functions and their interrelationships as the sufficient conditions for the viability of any human or social system (see [10]). These are applicable in a recursive mode, for example, to the different levels of an organization. The VSM has been widely applied in the diagnostic mode, but also to support the design of all kinds of social systems. Specific methodologies for these purposes have been developed, for instance for use in consultancy. The term viable system diagnosis (VSD) is also used.

The methodologies and models addressed up to this point have by and large been created in the positivistic tradition of science. Other strands in this tradition do exist, e. g., systems analysis and systems engineering, which together with OR have been called "hard systems thinking" (p. 127 in [31]). Also, more recent developments such as mathematical complexity and network theories, agent-based modeling and most versions of game theory can be classified as hard systems approaches.

The respective approaches have not altogether been excluded from fertile contacts with the interpretivist strand of inquiry. In principle, all of them can be considered as instruments for supporting discourses about different interpretations of an organizational reality or alternative futures studied in concrete cases. In our time, most applications of the VSM, for example, are constructivist in nature. To put it in a nutshell, these applications are (usually collective) constructions of a (new) reality, in which observation and interpretation play a crucial part. In this process, the actors involved make sense of the system under study, i. e., the organization in focus, by mapping it on the VSM. At the same time they bring forth "multiple realities rather than striving for a fit with one reality" (p. 299 in [29]).

The second group of methodologies is part of the interpretive strand:

- *Interactive Planning*. IP is a methodology, designed by Russell Ackoff [1], and developed further by Jamshid Gharajedaghi [28], for the purpose of dealing with "messes" and enabling actors to design their desired futures, as well as to bring them about. It is grounded in theoretical work on purposeful systems, reverts to the principles of continuous, participative and holistic planning, and centers on the idea of an "idealized design."
- *Soft Systems Methodology*. SSM is a heuristic designed by Peter Checkland [13,14] for dealing with complex situations. Checkland suggests a process of inquiry constituted by two aspects: A conceptual one, which is logic based, and a sociopolitical one, which is concerned with the cultural feasibility, desirability and implementation of change.
- *Critical Systems Heuristics*. CSH is a methodology, which Werner Ulrich [67,68] proposed for the purpose of scientifically informing planning and design in order to lead to an improvement in the human condition. The process aims at uncovering the interests that the system under study serves. The legitimacy and expertise of actors, and particularly the impacts of decisions and behaviors of the system on others – the "affected" – are elicited by means of a set of boundary questions. CSH can be seen as part of a wider move-

ment known as the "Emancipatory Systems Approach" which embraces, e. g., Freire's Critical Pedagogy, Interpretive Systemology, and Community OR (see pp. 291ff in [31]).

All three of these methodologies (IP, SSM, and CSH) are positioned in the interpretive tradition. Other methodologies and concepts which can be subsumed under the interpretive systems approach are, e. g., Warfield's science of generic design, Churchman's social system design, Senge's soft systems thinking, Mason and Mitroff's strategic assumptions surfacing and testing (SAST), Eden and Ackermann's strategic options in development and analysis (SODA), and other methodologies of soft operational research (for details, see pp. 211ff in [31]). The interpretive methodologies were designed to deal with qualitative aspects in the analysis and design of complex systems, emphasizing the communicational, social, political and ethical dimensions of problem solving. Several authors mention explicitly that they do not preclude the use of quantitative techniques or include such techniques in their repertoire (e. g., the biocyberneticist Frederic Vester).

In an advanced understanding of system dynamics both of these traditions—positivist and interpretivist—are synthesized. The adherents of SD conceive of model building and validation as a semi-formal, relativistic, holistic social process. Validity is understood as usefulness or fitness in relation to the purpose of the model, and validation as an elaborate set of procedures – including logico-structural, heuristic, algorithmic, statistical, and also discursive components – by which the quality of and the confidence in a model are gradually improved (see [4,5,59]).

System Dynamics – Its Features, Strengths and Limitations

The features, strengths and limitations of the SD methodology are a consequence of its specific characteristics. In the context of the multiple theories and methodologies of the systems movement, some of the distinctive features of SD are (for an overview, see [52], pp. 142ff in [31]):

- *Feedback as conceptual basis*. SD model systems are high-order, multiple-loop networks of closed loops of information. Concomitantly, an interest in non-linearities, long-term patterns and internal structure rather than external disturbances is characteristic of SD (p. 31 in [40]). However, SD models are not "closed systems", as sometimes is claimed, in the sense that (a) flows can originate from outside the system's boundaries, (b) representations of exogenous factors or systems can be incorporated into any model as parameters or special modules, and (c) new information can be accommodated via changes to a model. In other words, the SD view hinges on a view of systems which are closed in a causal sense but not materially (p. 297 in [52]).
- *Focus on internally generated dynamics*. SD models are conceived as closed systems. The interest of users is in the dynamics generated inside those systems. Given the nature of closed feedback loops and the fact that delays occur within them, the dynamic behavior of these systems is essentially non-linear.
- *Emphasis on understanding*. For system dynamicists the understanding of the dynamics of a system is the first goal to be achieved by means of modeling and simulation. Conceptually, they try to understand events as embedded in patterns of behavior, which in turn are generated by underlying structures. Such understanding is enabled by SD as it "shows how present policies lead to future consequences" (Sect. VIII in [23]). Thereby, the feedback loops are "a major source of puzzling behavior and policy difficulties" (p. 300 in [52]). SD models purport to test mental models, hone intuition and improve learning (see [65]).
- *High degree of operationality*. SD relies on formal modeling. This fosters disciplined thinking; assumptions, underlying equations and quantifications must be clarified. Feedback loops and delays are visualized and formalized; therewith the causal logic inherent in a model is made more transparent and discussable than in most other methodologies [53]. Also, a high level of realism in the models can be achieved. SD is therefore apt to support decision-making processes effectively.
- *Far-reaching requirements (and possibilities) for the combination of qualitative and quantitative aspects of modeling and simulation*. This is a consequence of the emphasis on understanding. The focus is not on point-precise prediction, but on the generation of insights into the patterns generated by the systems under study.
- *High level of generality and scale robustness*. The representation of dynamic systems in terms of stocks and flows is a generic form, which is adequate for a wide spectrum of potential applications. This spectrum is both broad as to the potential subjects under study, and deep as to the possible degrees of resolution and detail [38]. In addition, the SD methodology enables one to deal with large numbers of variables within multiple interacting feedback loops (p. 9 in [22]). SD has been applied to the most diverse subject areas, e. g., global modeling, environmental issues, social and economic policy, corporate and public management, regional planning, medicine, psychology and education in mathematics, physics and biology.

The features of SD just sketched out result in both strengths and limitations. We start with the strengths.

Strengths of SD

1. Its *specific modeling approach* makes SD particularly helpful in gaining insights into the patterns exhibited by dynamic systems, as well as the structures underlying them. Closed-loop modeling has been found most useful in fostering understanding of the dynamic functioning of complex systems. Such understanding is especially facilitated by the principle of modeling the systems or issues under study in a continuous mode and at rather high aggregation levels [20,38]. With the help of relatively small but insightful models, and by means of sensitivity analyses as well as optimization heuristics incorporated in the application software packages, decision-spaces can be thoroughly explored. Vulnerabilities and the consequences of different system designs can be examined with relative ease.
2. The *generality of the methodology* and its power to crystallize operational thinking in realistic models have triggered applications in the most varied contexts. Easy-to-use software and the features of screen-driven modeling via graphic user interfaces provide a strong lever for collaborative model-building in teams (cf. [2,69]).
3. Another strong point is the *momentum of the SD movement*. Due to the strengths commented above this point, the community of users has grown steadily, being probably the largest community within the systems movement. Lane (p. 484 in [36]) has termed SD "one of the most widely used systems approaches in the world."
4. Its specific features make SD an exceptionally effective tool for *conveying systemic thinking* to anybody. Therefore, it also has an outstanding track-record of classroom applications for which "learner-directed learning" [24] or "learner-centered learning" is advocated [25,26]. Pertinent audiences range from schoolchildren at the levels of secondary and primary schools to managers and scientists.

Given these strengths, the community of users has not only grown significantly, but has also transcended disciplinary boundaries, ranging from the formal and natural sciences to the humanities, and covering multiple uses from theory building and education to the tackling of real-world problems at almost any conceivable level. Applications to organizational, societal and ecological issues have seen a particularly strong growth. This feeds back on the availability and growth of the knowledge upon which the individual modeler can draw.

The flip side of most of the strengths outlined here embodies the limitations of SD; we concentrate on those which can be relevant to a possible complementarity of SD with other systems methodologies.

Limitations of SD

1. The main point here is that SD does not provide a framework or methodology for the *diagnosis and design* of organizational structures in the sense of interrelationships among organizational actors. This makes SD susceptible to completion from without – a completion which organizational cybernetics (OC), and the VSM in particular, but also living system theory (LST), especially can provide. The choice falls on these two approaches because of their strong heuristic power and their complementary strengths in relation to SD (cf. [57,61]).
2. Another limitation of SD is related to the *absorption of variety* (complexity) by an organization. *Variety* is a technical term for *complexity*, which denotes a (high) number of potential states or behaviors of a system (based on [3,8]). SD offers an approach to the handling of variety which allows modeling at different scales of a problem or system [47]. It focuses on the identification, at a certain resolution level or possibly several resolution levels, of the main stock variables which will be affected by the respective flows. These, in turn, will be influenced by parameters and auxiliary variables. This approach, even though it enables thinking and modeling at different scales, does not provide a formal procedure for an organization to cope with the external complexity it faces, namely, for designing a structure which can absorb that complexity. In contrast, OC and LST offer elaborate models to enable the absorption of variety, in the case of the VSM based explicitly on Ashby's *Law of Requisite Variety*. It says "Only variety can destroy variety", which implies that the varieties of two interacting systems must be in balance, if stability is to be achieved [3]. The VSM has two salient features in this respect. Firstly, it helps design an organizational unit for viability, by enabling it to attenuate the complexity of its environment, and also to enhance its eigenvariety, so that the two are in balance. The term *variety engineering* has been used in this context [9]. Secondly, the recursive structure of the VSM ensures that an organization with several levels will develop sufficient eigenvariety along the fronts on which the complexity it faces unfolds. Similarly, LST offers the conditions for social systems to survive, by maintaining thermodynamically highly improbable energy states via continuous inter-

action with their environments. The difference between the two approaches is that the VSM functions more in the strategic and informational domains, while the LST model essentially focuses on the operational domain. In sum, both can make a strong contribution related to coping with the external complexity faced by organizations, and therefore can deliver a strong complement to SD.

3. Finally, the design of *modeling processes* confronts SD with specific challenges. The original SD methodology of modeling and simulation was to a large extent functionally and technically oriented. This made it strong in the domain of logical analysis, while the socio-cultural and political dimensions of the modeling process were, if not completely out of consideration, at least not a significant concern in methodological developments. The SD community – also under the influence of the soft systems approaches – has become aware of this limitation and has worked on incorporating features of the social sciences into its repertoire. The following examples, which document this effort to close the gap, stand for many. Extensive work on group model building has been achieved, which explores the potential of collaborative model building [69]. A new schema for the modeling process has been proposed, which complements logic-based analysis by cultural analysis [37]. The social dimension of system dynamics-based modeling has become subject to intensive discussion ([77]; and other contributions to the special issue of *Systems Research and Behavioral Science*, Vol. 51, No. 4, 2006). Finally, in relation to consultancy methodology, modeling has been framed as a learning process [34] and as second-order intervention [60].

As has been shown, there is a need to complement classical SD with other methodologies, when issues are at stake which it cannot handle by itself. VSM and LST are excellent choices when issues of organizational diagnosis or design are to be tackled.

The limitations addressed here call attention to other methodologies which exhibit certain features that traditionally were not incorporated, or at least not explicit, in SD methodology. One aspect concerns the features that explicitly address the subjectivity of purposes and meanings ascribed to systems. In this context, support for problem formulation, model construction and strategy design by individuals on the one hand and groups on the other are relevant issues. Also, techniques for an enhancement of creativity (e. g., the generation and the reframing of options) in both individuals and groups are a matter of concern. Two further aspects relate to methodological arrangements for coping with the specific issues of negotiation and alignment in pluralist and coercive settings.

As far as the modeling processes are concerned, group model building has proven to be a valuable complement to pure modeling and simulation. However, there are other systems methodologies which should be considered as potentially apt to enrich SD analysis, namely the soft approaches commented upon earlier, e. g., interactive planning, soft system methodology and critical system heuristics.

On the other hand, SD can be a powerful complement to other methodologies which are more abstract or more static in nature. This potential refers essentially to all systems approaches which stand in the interpretive (“soft”) tradition, but also to approaches which stand in the positivist traditions, such as the VSM and LST. These should revert to the support of SD in the event that tradeoffs between different goals must be handled, or if implications of long-term decisions on short-term outcomes (and vice versa) have to be ascertained, and whenever contingencies or vulnerabilities must be assessed.

Actual and Potential Relationships

It should be clear by now that the systems movement has bred a number of theories and methodologies, none of which can be considered all-embracing or complete. All of them have their strengths and weaknesses, and their specific potentials and limitations.

Since Burrell and Morgan [12] adverted to incommensurability between different paradigms of social theory, several authors have acknowledged or even advocated methodological complementarism. They argue that there is a potential complementarity between different methods, and, one may add, models, even if they come from distinct paradigms. Among these authors are, e. g., Brocklesby [11], Jackson [30], Midgley [43], Mingers [45], Schwaninger [55] and Yolles [83]. These authors have opened up a new perspective in comparison with the non-complementaristic state-of-the-art.

In the past, the different methodologies have led to the formation of their own traditions and “schools,” with boundaries across which not much dialogue has evolved. The methodologies have kept their protagonists busy testing them and developing them further. Also, the differences between different language games and epistemological traditions have often suggested incommensurability, and therewith have impaired communication. Prejudices and a lack of knowledge of the respective other side have accentuated this problem: Typically, “hard” systems scientists are suspicious of "soft" systems scientists. For example, many members of the OR community, not unlike orthodox quantitatively oriented economists, adhere to the opinion that "SD is too soft." On the other hand the protagonists of "soft" systems approaches, even though many of them have adopted feedback diagrams (causal loop diagrams) for the sake of visualization, are all too often convinced that "SD is too hard." Both of these judgments indicate a lack of knowledge, in particular of the SD validation and testing methods available, on the one hand, and the technical advancements achieved in modeling and simulation, on the other (see [5,59,66]).

In principle, both approaches are complementary. The qualitative view can enrich quantitative models, and it is connected to their philosophical, ethical and esthetical foundations. However, qualitative reasoning tends to be misleading if applied to causal network structures without being complemented by formalization and quantification of relationships and variables. Furthermore, the quantitative simulation fosters insights into qualitative patterns and principles. It is thus a most valuable device for validating and honing the intuition of decision makers, via corroboration and falsification.

Proposals that advocate mutual learning between the different "schools" have been formulated inside the SD community (e. g., [35]). The International System Dynamics Conference of 1994 in Stirling, held under the banner of "Transcending the Boundaries," was dedicated to the dialogue between different streams of the systems movement.

Also, from the 1990s onwards, there were vigorous efforts to deal with methodological challenges, which traditionally had not been an important matter of scientific interest within the SD community. Some of the progress made in these areas is documented in a special edition of *Systems Research and Behavioral Science* (Vol. 21, No. 4, July-August 2004). The main point is that much of the available potential is based on the complementarity, not the mutual exclusiveness, of the different systems approaches.

In the future, much can be gained from leveraging these complementarities. Here are two examples of methodological developments in this direction, which appear to be achievable and potentially fertile: The enhancement of qualitative components in "soft" systems methodologies in the process of knowledge elicitation and model building (cf. [69]), and the combination of cybernetics-based organizational design with SD-based modeling and simulation (cf. [61]). Potential complementarities exist not only across the qualities – quantities boundary, but also within each one of the domains. For example, with the help of advanced software, SD modeling ("top-down") and agent-based modeling ("bottom-up") can be used in combination.

From a meta-methodological stance, generalist frameworks have been elaborated which contain blueprints for combining different methodologies where this is indicated. Two examples are:

- *Total systems intervention* (TSI) is a framework proposed by Flood and Jackson [19], which furnishes a number of heuristic schemes and principles for the purpose of selecting and combining systems methods/methodologies in a customized way, according to the issue to be tackled. SD is among the recommended "tools".
- *Integrative systems methodology* (ISM) is a heuristic for providing actors in organizations with requisite variety, developed by Schwaninger [55,56]. It advocates (a) dealing with both content– and context-related issues during the process, and (b) placing a stronger emphasis on the validation of qualitative and quantitative models as well as strategies, in both dimensions of the content of the issue under study and the organizational context into which that issue is embedded. For this purpose, the tools of SD (to model content) and organizational cybernetics – the VSM (to model context) – are cogently integrated.

These are only two examples. In principle, SD could make an important contribution in the context of most of the methodological frameworks, far beyond the extent to which this has been the case. Systems methodologists and practitioners can potentially benefit enormously from including SD methodology in their repertoires.

Outlook

There have recently been calls for an eclectic "mixing and matching" of methodologies. In light of the epistemological tendencies of our time towards radical relativism, it is necessary to warn against taking a course in which "anything goes". It is most important to emphasize that the desirable methodological progress can only be achieved on the grounds of scientific rigor. This postulate of "rigor" is not to be confused with an encouragement of "rigidity." The necessary methodological principles advocated here are disciplined thinking, a permanent quest for better models (that is, thorough validation), and the highest achievable levels of transparency in the formalizations as well as of the underlying assumptions and sources used. Scientific rigor, in this context, also implies that combinations of methodologies reach beyond merely eclectic addons from different methodologies, so that genuine integration towards better adequacy to the issues at hand is achieved.

The contribution of system dynamics can come in the realms of the following:

- Fostering disciplined thinking
- Understanding dynamic behaviors of systems and the structures that generate them
- Exploring paths into the future and the concrete implications of decisions
- Assessing strategies as to their robustness and vulnerabilities, in ways precluded by other, more philosophical, and generally "soft" systems approaches

These latter streams can contribute to reflecting and tackling the meaning- and value-laden dimensions of complex human, social and ecological systems. Some of their features should and can be combined synergistically with system dynamics, particularly by being incorporated into the repertoires of system dynamicists. From the reverse perspective, incorporating system dynamics as a standard tool will be of great benefit for the broad methodological frameworks. Model formalization and dynamic simulation may even be considered necessary components for the study of the concrete dynamics of complex systems.

Finally, there are also many developments in the "hard", i. e., mathematics-, statistics-, logic-, and informatics-based methods and technologies, which are apt to enrich the system dynamics methodology, namely in terms of modeling and decision support. For example, the constantly evolving techniques of time-series analysis, filtering, neural networks and control theory can improve the design of system-dynamics-based systems of (self-)control. Also, a bridge across the divide between the top-down modeling approach of SD and the bottom-up approach of agent-based modeling appears to be feasible. Furthermore, a promising perspective for the design of genuinely "intelligent organizations" emerges if one combines SD with advanced database-management, cooperative model building software, and the qualitative features of the "soft" systems methodologies.

The approaches of integrating complementary methodologies outlined in this contribution definitely mark a new phase in the history of the systems movement.

Appendix

Milestones in the Evolution of the Systems Approach in General and System Dynamics in Particular

The table gives an overview of the systems movement's evolution, as shown in its main literature; and that overview is not exhaustive.

Systems Approaches – An Overview

Note: This diagram shows three streams of the systems approach in the context of their antecedents. The general systems thread has its origins in philosophical roots from antiquity: The term *system* derives from the old Greek *σύστημα* (systēma), while, *cybernetics* stems from the Greek *κυβερνήτης* (kybernētēs). The arrows between the threads stand for interrelationships and efforts to synthesize the connected approaches. For example, integrated systems methodology is an integrative attempt to leverage the complementarities of system dynamics and organizational cybernetics. Enumerated to the left and right of the scheme are the fields of application. The big arrows in the upper region of the diagram indicate that the roots of the systems approach continue influencing the different threads and the fields of application even if the path via general systems theory is not pursued.

The diagram is not a complete representation, but the result of an attempt to map the major threads of the systems movement and some of their interrelations. Hence, the schema does not cover all schools or protagonists of the movement. Why does the diagram show a dynamic and evolutionary systems thread and a cybernetics thread, if cybernetics is about dynamic systems? The latter embraces all the approaches that are explicitly grounded in cybernetics. The former relates to all other approaches concerned with dynamic or evolutionary systems. The simplification made it necessary to somewhat curtail logical perfection for the sake of conveying a synoptic view of the different systems approaches, in a language that uses the categories common in current scientific and professional discourse. Overlaps exist, e. g., between dynamic systems and chaos theory, cellular automata and agent-based modeling.

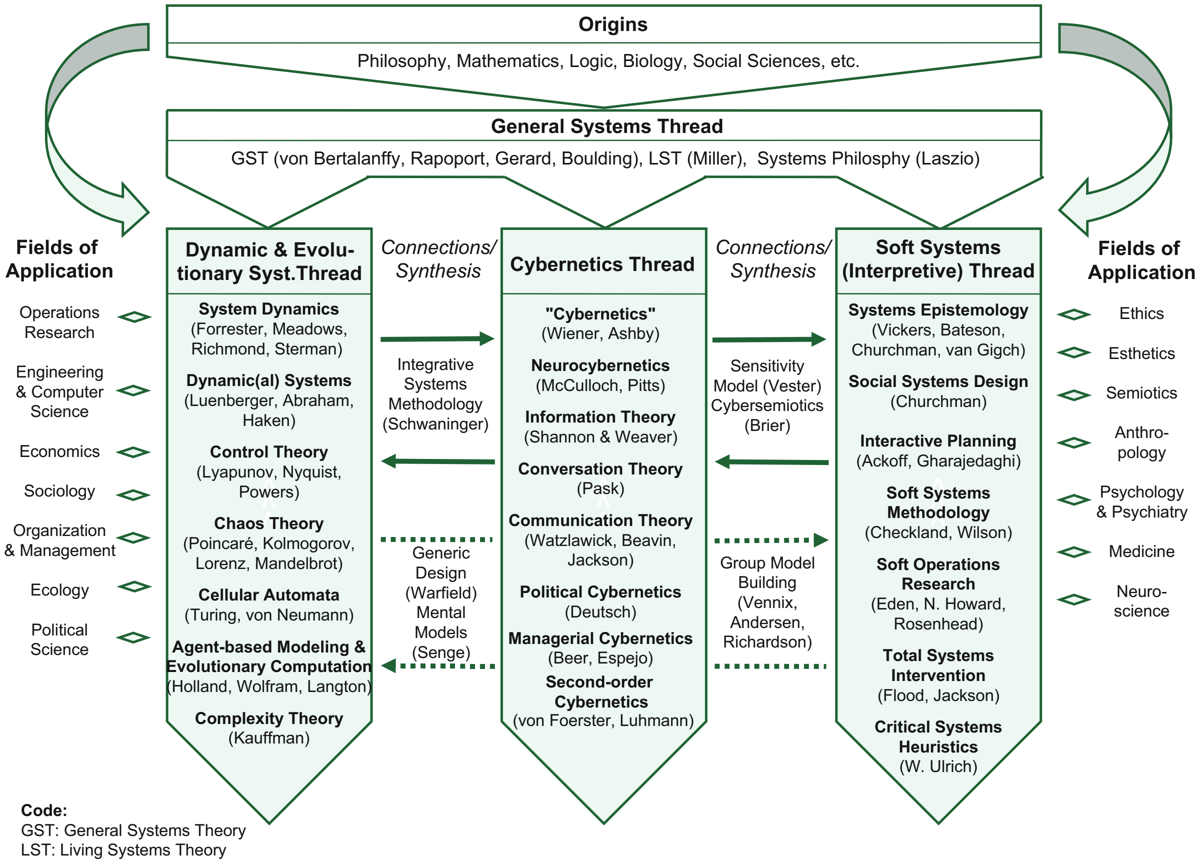
Bibliography

Primary Literature

1. Ackoff RL (1981) Creating the Corporate Future. Wiley, New York
2. Andersen DF, Richardson GP (1997) Scripts for Group Model Building. Syst Dyn Rev 13(2):107–129
3. Ashby WR (1956) An Introduction to Cybernetics. Chapman & Hall, London
4. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
5. Barlas Y, Carpenter S (1990) Philosophical roots of model validity: Two paradigms. Syst Dyn Rev 6(2):148–166
6. Beer S (1959) Cybernetics and Management. English Universities Press, London
7. Beer S (1994) Towards the Cybernetic Factory. In: Harnden R, Leonard A (eds) How Many Grapes Went into the Wine. Stafford Beer on the Art and Science of Holistic Management.

System Dynamics in the Evolution of the Systems Approach, Table 1
Milestones in the evolution of the systems approach in general and system dynamics in particular

Foundations of general system theory		
Von Bertalanffy	Zu einer allgemeinen Systemlehre	1945
	An Outline of General System Theory	1950
	General System Theory	1968
Bertalanffy, Boulding, Gerard, Rapoport	Foundation of the Society for General Systems Research	1953
Klir	An Approach to General System Theory	1968
Simon	The Sciences of the Artificial	1969
Pichler	Mathematische Systemtheorie	1975
Miller	Living Systems	1978
Mesarovic & Takahara	Abstract Systems Theory	1985
Rapoport	General System Theory	1986
Foundations of cybernetics		
Macy Conferences (Josiah Macy, Jr. Foundation)	Cybernetics. Circular Causal, and Feedback Mechanisms in Biological and Social Systems	1946–1951
Wiener	Cybernetics or Control and Communication in the Animal and in the Machine	1948
Ashby	An Introduction to Cybernetics	1956
Pask	An Approach to Cybernetics	1961
Von Foerster, Zopf	Principles of Self-Organization	1962
McCulloch	Embodiments of Mind	1965
Foundations of organizational cybernetics		
Beer	Cybernetics and Management	1959
	Towards the Cybernetic Factory	1962
	Decision and Control	1966
	Brain of the Firm	1972
Von Foerster	Cybernetics of Cybernetics	1974
Foundations of system dynamics		
Forrester	Industrial Dynamics	1961
	Principles of Systems	1968
	Urban Dynamics	1969
	World Dynamics	1971
Meadows et al.	Limits to Growth	1972
Richardson	Feedback Thought in Social Science and Systems Theory	1991
Systems methodology		
Churchman	Challenge to Reason	1968
	The Systems Approach	1968
Vester & von Hesler	Sensitivitätsmodell	1980
Checkland	Systems Thinking, Systems Practice	1981
Ackoff	Creating the Corporate Future	1981
Ulrich	Critical Heuristics of Social Planning	1983
Warfield	A Science of Generic Design	1994
Schwaninger	Integrative Systems Methodology	1997
Gharajedaghi	Systems Thinking	1999
Sabelli	Bios – A Study of Creation	2005
Selected recent works in system dynamics		
Senge	The Fifth Discipline	1990
Barlas & Carpenter	Model Validity	1990
Vennix	Group Model Building	1996
Lane & Oliva	Synthesis of System Dynamics and Soft Systems Methodology	1998
Sterman	Business Dynamics	2000
Warren	Strategy Dynamics	2002, 2008
Wolstenholme	Archetypal Structures	2003
Morecroft	Strategic Modelling	2007
Schwaninger & Grösser	Theory-building with System Dynamics & Model Validation	2008, 2009



System Dynamics in the Evolution of the Systems Approach, Figure 1

Wiley, Chichester, pp 163–225 (reprint, originally published in 1962)

8. Beer S (1966) Decision and Control. Wiley, Chichester
9. Beer S (1979) The Heart of Enterprise. Wiley, Chichester
10. Beer S (1981) Brain of the Firm, 2nd edn. Wiley, Chichester
11. Brocklesby J (1993) Methodological complementarism or separate paradigm development – Examining the options for enhanced operational research. Aust J Manag 18(2):133–157
12. Burrell G, Morgan G (1979) Sociological Paradigms and Organisational Analysis. Hants, Gower
13. Checkland PB (1981) Systems Thinking, Systems Practice. Wiley, Chichester
14. Checkland PB, Poulter J (2006) Learning for Action: A Short Definitive Account of Soft Systems Methodology, and its Use Practitioners, Teachers and Students. Wiley, Chichester
15. Churchman CW (1968) Challenge to Reason. McGraw-Hill, New York
16. Churchman CW (1968) The Systems Approach. Delacorte Press, New York
17. Churchman CW (1979) The Systems Approach and its Enemies. Basic Books, New York
18. Encyclopedia of Life Support Systems (2002) published under: http://www.eolss.net/
19. Flood RL, Jackson MC (1991) Creative Problem Solving. Total Systems Intervention. Wiley, Chichester
20. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
21. Forrester JW (1968) Principles of Systems. MIT Press, Cambridge
22. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
23. Forrester JW (1971) World Dynamics. Pegasus Communications, Waltham
24. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policy Making. Kluwer, Boston
25. Forrester JW (1993) System Dynamics as an Organizing Framework for Pre-college Education. Syst Dyn Rev 9(2):183–194
26. Forrester JW (1997) System Dynamics and K-12 Teachers. A Lecture at the University of Virginia School of Education, Massachusetts Institute of Technology. System Dynamics Group Paper D-4665–4
27. François C (2004) International Encyclopedia of Systems and Cybernetics, 2nd edn. Saur, München
28. Gharajedaghi J (1999) Systems Thinking. Managing Chaos and Complexity. Butterworth-Heinemann, Boston
29. Harnden RJ (1989) Technology for Enabling: The Implications for Management Science of a Hermeneutics of Distinction. The University of Aston, Birmingham
30. Jackson MC (1991) Systems Methodology for the Management Sciences. Plenum Press, New York

31. Jackson MC (2000) Systems Approaches to Management. Kluwer Academic/Plenum, New York
32. Kauffman SA (1993) The Origins of Order. Self-Organization and Selection in Evolution. Oxford University Press, New York
33. Klir GJ (1969) An Approach to General Systems Theory. Nostrand, New York
34. Lane DC (1994) Modeling as Learning: A Consultancy Methodology for Enhancing Learning in Management in Management Teams. In: Morecroft J, Sterman JD (eds) Modeling for Learning Organizations. Productivity Press, Portland, pp 205–240
35. Lane DC (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10(2–3):101–134
36. Lane DC (2006) IFORS' Operational Research Hall of Fame. Jay Wright Forrester. Int Trans Oper Res 13:483–492
37. Lane DC, Oliva R (1998) The Greater Whole: towards a synthesis of system dynamics and soft system methodology. Eur J Oper Res 107(1):214–235
38. La Roche U, Simon M (2000) Geschäftsprozesse simulieren: flexibel und zielorientiert führen mit Fliessmodellen. Orell Füssli, Zürich
39. McCulloch WS (1965) Embodiments of Mind. MIT Press, Cambridge
40. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the System Dynamics Method. MIT Press, Cambridge, pp 23–57
41. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) Limits to Growth. Universe Books, New York
42. Mesarovic MD, Takahara Y (1985) Abstract Systems Theory. Springer, Berlin
43. Midgley G (2000) Systemic Intervention. Philosophy, Methodology, and Practice. Kluwer, New York
44. Miller JG (1978) Living Systems. McGraw-Hill, New York
45. Mingers J (1997) Multi-paradigm Multimethodology. In: Mingers J, Gill A (eds) Multimethodology. Wiley, Chichester
46. Morecroft J (2007) Strategic Modelling and Business Dynamics: a Feedback Systems Approach. Wiley, Chichester
47. Odum HT, Odum EC (2000) Modeling for all Scales: An Introduction to System Simulation. Academic Press, San Diego
48. Pask G (1961) An Approach to Cybernetics. Hutchinson, London
49. Pichler F (1975) Mathematische Systemtheorie. de Gruyter, Berlin
50. Rapoport A (1953) Operational philosophy: Integrating Knowledge and Action. Harper, New York
51. Rapoport A (1986) General System Theory. Essential Concepts and Applications Abacus Press, Turnbridge Wells
52. Richardson GP (1999) Feedback Thought in Social Science and Systems Theory. Pegasus Communications, Waltham (Originally published in 1991)
53. Richmond B (1997) The "Thinking" in systems thinking: How can we make it easier to master? Syst Think 8(2):1–5
54. Sabelli H (2005) Bios: a Study of Creation. World Scientific, Hackensack
55. Schwaninger M (1997) Integrative systems methodology: Heuristic for requisite variety. Int Trans Oper Res 4(4):109–123
56. Schwaninger M (2004) Methodologies in conflict: Achieving synergies between system dynamics and organizational cybernetics. Syst Res Behav Sci 21(4):1–21
57. Schwaninger M (2006) Theories of viability. A comparison. Syst Res Behav Sci 23:337–347
58. Schwaninger M, Groesser S (2008) System dynamics as model-based theory building. Syst Res Behav Sci 25:1–19
59. Schwaninger M, Groesser S (2009) Model Validation: The Quest for Quality in System Dynamics Modeling. Encyclopaedia of Complexity and Systems Science. Springer, New York
60. Schwaninger M, Janovjak M, Ambroz K (2006) Second-order intervention: Enhancing organizational competence and performance. Syst Res Behav Sci 23:529–545
61. Schwaninger M, Pérez Ríos J (2008) System dynamics and cybernetics: A synergetic pair. Syst Dyn Rev 24(2):145–174
62. Senge PM (1990) The Fifth Discipline. The Art and Practice of the Learning Organization. Doubleday, New York
63. Shapiro M, Mandel T, Schwaninger M et al (1996) The Primer Toolbox. International Society for the Systems Sciences, http://www.isss.org/primer/toolbox.htm
64. Simon HA (1969) The Sciences of the Artificial. MIT Press, Cambridge
65. Sterman JD (1994) Learning in and about complex systems. Syst Dyn Rev 10(2–3):291–330
66. Sterman JD (2000) Business Dynamics. Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
67. Ulrich W (1983) Critical Heuristics of Social Planning. Haupt, Bern
68. Ulrich W (1996) A Primer to Critical Systems Heuristics for Action Researchers. The Centre of Systems Studies, University of Hull, Hull
69. Vennix JAM (1996) Group Model Building. Facilitating Team Learning Using System Dynamics. Wiley, Chichester
70. Vester F, Von Hesler A (1980) Sensitivitätsmodell. Regionale Planungsgemeinschaft Untermain, Frankfurt am Main
71. Von Bertalanffy L (1949) Zu einer allgemeinen Systemlehre. Bl Dtsch Philos 18(3/4) (Excerpts: in Biol Gen 19(1):114–129 and in General System Theory, 1968, Chapter III)
72. Von Bertalanffy L (1950) An outline of general system theory. Br J Philos Sci 1:139–164
73. Von Bertalanffy L (1968) General System Theory. Braziller, New York
74. Von Foerster H (1984) Observing Systems, 2nd edn. Intersystems Publications, Seaside
75. Von Foerster H (ed) (1995) Cybernetics of Cybernetics, 2nd edn. Future Systems, Minneapolis (Originally published in 1974)
76. Von Foerster H, Zopf GW (eds) (1962) Principles of Self-Organization. Pergamon Press, Oxford
77. Vriens D, Achtenbergh J (2006) The social dimension of system dynamics-based modelling. Syst Res Behav Sci 23(4):553–563
78. Warfield JN (1994) A Science of Generic Design: Managing Complexity through Systems Design, 2nd edn. Iowa State University Press, Ames
79. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
80. Warren K (2008) Strategic Management Dynamics. Wiley, Chichester
81. Wiener N (1948) Cybernetics: Control and Communication in the Animal and in the Machine. MIT Press, Cambridge
82. Wolstenholme E (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

83. Yolles MA (1998) Cybernetic exploration of methodological complement. Kybern 27(5):527–542

Books and Reviews

Jackson MC (2003) Systems Thinking: Creative Holism for Managers. Wiley, Chichester

Klir GJ (2001) Facets of Systems Science, 2nd edn. Kluwer Academic/Plenum, New York

Midgley G (ed) (2003) Systems Thinking, vol 4. Sage, London

Ragsdell G, Wilby J (eds) (2001) Understanding Complexity. Kluwer Academic/Plenum, New York

Richardson GP (ed) (1996) Modelling for Management. Simulation in Support of Systems Thinking, 2 Volumes. Aldershot, Dartmouth

Schwaninger M (2006) Intelligent Organizations. Powerful Model for Systemic Management. Springer, Berlin

Van Gigch JP (2003) Metadecisions. Rehabilitating Epistemology. Kluwer Academic/Plenum, New York

System Dynamics Modeling: Validation for Quality Assurance

MARKUS SCHWANINGER, STEFAN GROESSER
Institute of Management, University of St. Gallen,
St. Gallen, Switzerland

Article Outline

Glossary

Model/model system A model is a simplified representation of a real system. Models can be descriptive or prescriptive (normative). Their functions can be to enable explanation, anticipation or design. A distinction used in this contribution is between causal and non-causal models, with System Dynamics models being of the former type. The term *model system* is used to stress the systemic character of a model; this serves to identify it as an organized whole of variables and relationships on the one hand, and to distinguish it from the *real system* which is to be modeled, on the other.

Model validity A model's property of adequately reflecting the system modeled. Validity is the primary measure of model quality. It is a matter of degree, not a dichotomized property.

Model purpose The goal for which a model is designed or the function it is intended to fulfill. The model purpose is closely linked to the end-model user or model owner. Model purpose is the criterion for the choice of a model's boundary and design.

Modeling process The process involving phases such as problem articulation, boundary selection, development of a dynamic hypothesis, model formulation, model testing, policy formulation and policy evaluation [28]. The modeling process is followed by model use and implementation, i. e., the realization of actions designed or facilitated by the use of the model.

Validation process Validation is the process by which model validity is enhanced systematically. It consists in gradually building confidence in the usefulness of a model by applying validation tests as outlined in this chapter. In principle, validation pervades all phases of the modeling process, and, in addition, extends into the phases of model use and implementation.

Definition of the Subject

The present chapter addresses the question of building better models. This is crucial for coping with complexity in general, and in particular for the management of dynamic systems. Both the epistemological and the methodological-technological aspects of model validation for the achievement of high-quality models are discussed. The focus is on formal models, i. e. those formulated in a stringent, logical, and mostly mathematical language.

Introduction

The etymological root of *valid* is the Latin word *validus*, which denotes attributes such as strong, powerful and firm. A valid model, then, is well-founded and difficult to reject because it accurately represents the perceived real system which it is supposed to reflect. This system can be either one that already exists or one that is being constructed, or even anticipated, by a modeler or a group of modelers.

Validation standards in System Dynamics are more rigorous than those of many other methodologies. Let us distinguish between two types of mathematical models, which are fundamentally different: Causal, theory-like models and non-causal, statistical (correlational) models [4]. The former are explanatory, i. e., they embody theory about the functioning of a real system. The latter are descriptive and express observed associations among different elements of a real system. System Dynamics models are causal models.

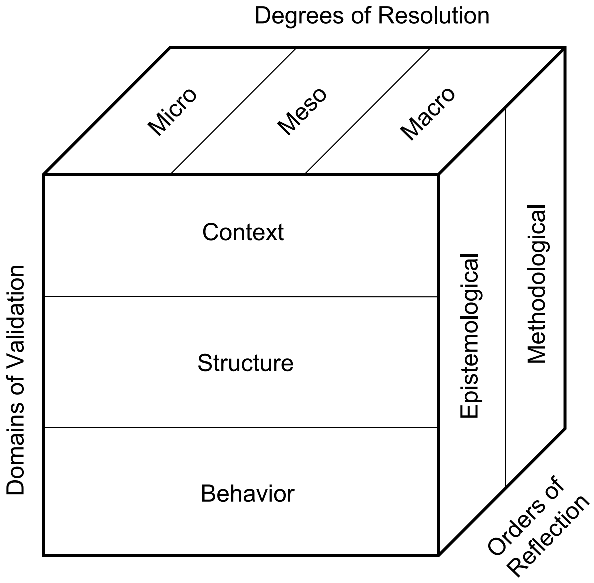
Non-causal models are tested globally, in that the statistical fit between model and data series from the real system under study is assessed. If the fit is satisfactory, the model is considered to be accurate ("valid", "true"). In contrast, system dynamicists postulate that models be not only right, but right for the right reasons. As the models are made up of causal interdependencies, accuracy is required for each and every variable and relationship. The following principle applies: if only one component of the model is shown to be wrong, the whole model is rejected even if the overall model output fits the data [4]. This strict standard is conducive to high-quality modeling practice.

A model is an abstract version of a perceived reality. Simulation is a way of experimenting with mathematical models to gain insights and to employ these to improve the real system under study. It is often said that System Dynamics models should portray problems or issues, not systems. This statement must be interpreted in the sense that one should not try to set the boundaries of the model too widely, but rather give the model a focus by concentrating on an object in accordance with the specific purpose of the model. In a narrower definition, even an issue or problem can be conceived of as a "system", i. e., "a portion of the world sufficiently well defined to be the subject of study" [21]. Validity then consists in a stringent correspondence between model system and real system.

We will treat the issue of model validation as a means of assuring high-quality models. We interject that validity is not the only criterion of model quality, other criteria including parsimony, ease-of-use, practicality, importance, etc. [22].

In the following, the epistemological foundations of model validity are reviewed (Sect. "Epistemological Foundations"). Then, an overview of the methods for assuring model validity is given (Sect. "Validation Methods"). Further, the survey includes an overview of the validation process (Sect. "Validation Process") and our final conclusions (Sect. "Synopsis and Outlook").

The substance of this article will be made more palpable by means of the following frame of reference. We call it the Validation Cube. The diagram in Fig. 1 shows three dimensions of the validation topic:



System Dynamics Modeling: Validation for Quality Assurance, Figure 1
The Validation Cube – A frame of reference showing three dimensions of the validation topic

- *Orders of Reflection:* We distinguish between an *epistemological* and a *methodological* layer. These define the objects of the next two Sects. "Epistemological Foundations" and "Validation Methods".
- *Domains of Validation:* The three domains, *context*, *structure* and *behavior* refer to the groups of validation methods as described in Sect. "Validation Methods".
- *Degrees of Resolution:* We address the different granularities of models. *Micro* refers to the smallest building blocks of models (e. g., variables or small sets of variables), *meso* to modules which constitute a model, and *macro* to the model as a whole.

Epistemological Foundations

Epistemology is the theory that enquires into the nature and grounds of knowledge: "What can we know and how do we know it?" [13]. These questions are of utmost importance when dealing with models and their validity, because a method of validation is only as good as its epistemological basis.

We can only briefly refer to the antecedents of the epistemological perspective inherent in the idea of model validation as commonly held today in the community of system dynamicists. One could go back to Socrates who, in Plato's *Republic* (fourth century BC), addressed the problematic relationship between reality, image and knowledge. One could also refer to John Locke (seventeenth century), the first British empiricist who maintained that ideas could come only from experience, while admitting that our knowledge about external objects is uncertain. We will address the philosophical movements of the nineteenth and twentieth centuries, which are direct sources of the epistemology which is important for model validation. The reader may kindly excuse us for certain massive simplifications that we are obliged to make.

What will be said here about theories applies equally to formal models. In System Dynamics, models either embody theories or they are considered essential components of theories. In addition, processes of modeling and theory-building are of the same nature; a model, like any theory, is built and improved in a dialectic of propositions and refutations [22].

Positivism and Critique

Positivism is a scientific doctrine founded by Auguste Comte (nineteenth century) which raises the *positive* to the principle of all scientific knowledge. "Positive", in this context, is not meant to be the opposite of negative, but the given, factual, or indubitably existent. The positive is associated with features such as being real, useful, certain, and precise. Positivism confines science to the observable and manipulable, drawing on the mathematical, empirical orientation of the natural sciences as its paragon. The objectivist claim of positivism is that things exist independently of the mind and that truths are detached from human values and beliefs. This stance calls for models that approximate an objective reality.

A younger development in this vein is the school of logical positivism, also logical empiricism (with Schlick, Neurath, Hempel, etc.), which concentrates on the problem of meaning and has developed the verifiability principle: Something is meaningful only if verifiable empirically, i. e., ultimately by observation through the senses. To verify here means to show to be true [13]. For the logical positivists, the method of verification is the essence of theory-building. Tests of theories hinge on their confirmation by facts. In System Dynamics, testing models on real-world data is a core component of validation.

Positivism has been criticized for being reductionist, i. e., for its tendency to reduce concepts to simpler or empirically more accessible ones, and to conceive of learning as an accumulation of particular details. The critique has also asserted that there is no theory-independent identification of facts, and therefore different theories cannot be tested by means of the same data [6]. Another objection maintains that social facts are not merely given, but produced by human action, and that they are subject to interpretation [23]. These arguments introduce the principle of relativity, which is of crucial importance for the field of model validation: A model is a subjective construction by an observer.

Pragmatism – A Challenge to Positivism

Pragmatism, which arose in the second half of the nineteenth century, emphasizes action and the practical consequences of thinking. Its founder, Charles Sanders Peirce, was interested in the effects that the meaning of scientific concepts could have on human experience and action. He defined truth as "the opinion which is fated to be ultimately agreed to by all who investigate" [13], whereby truth is linked to consensual validation. For pragmatists, truth is in what works (Ferdinand Schiller) or satisfies us (John Dewey), and what we find believable and consistent: " 'The true' … is only the expedient in the way of our thinking", and "truth is *made* … in the course of experience." (see p. 581 and p. 583 in [11]).

Pragmatism is often erroneously disdained for supposedly being a crass variety of utilitarianism and embodying a crude instrumentalist rationality. A more accurate view considers the fact that pragmatists are not satisfied with a mere ascertainment of truth; instead they ask: "If an idea or assumption is true, does this make a concrete difference to the life of people? How can this truth be actualized?" In other words, pragmatism does not crudely equate truth and utility. It rather postulates that those truths which are useful to people ought to be put into practice [23].

Pragmatism introduces the criteria of confidence and usefulness, which are more operational as guides to the evaluation of experiments than is the notion of an absolute truth, which is unattainable in the realm of human affairs. At the same time, pragmatism triggers a crucial insight for the context of model-building: The validity of a model depends not only on the absolute quality of that model but also hinges on its suitability with respect to a purpose [7]. In the context of model validation, then, truth is a relative property; more exactly, a *truth* holds for a limited domain only.

More Challenges to Positivism

We discuss three more challenges to positivism in the twentieth century. First, Thomas Kuhn's theory of scientific revolutions [12]: Kuhn shows, by means of historical cases, that in the sphere of science, generally accepted ways of looking at the world ("paradigms") change over time through fundamental shifts. Therefore, the activities of a scientist are largely shaped by the dominant scientific worldview. Second, Willard Van Orman Quine and Wilfrid Sellars argue that knowledge creation and theory-building is a holistic, conversational process, as opposed to the reductionist and confrontational views [4].

Both of these movements contribute to our understanding of how real systems are to be modeled and validated: as organized wholes, and consciously with respect to the values and beliefs underlying a given modeling process. This approach adheres to the spirit of models themselves, by means of which the behavior of whole systems can be simulated and tested on their inherent assumptions.

A third challenge is presented by the interpretive streams of epistemology (for an overview, see [9]). Among them, a main force which expands the possibilities of scientific methodologies is the strand of hermeneutics. Derived from the Greek *hermeneuein* – to interpret or to explain – the term *hermeneutics* stands for a school, mainly associated with Hans-Georg Gadamer, which pursues the ideal of a human science of understanding. The emphasis is on interpretation in an interplay between a subject-matter and the interpreter's position. This emphasis introduces the subjective into scientific methodology. Hermeneutics denies both that a single "objective true interpretation" can transcend all individual viewpoints, and that humans are forever confined within their own ken [13]. This epistemology offers a necessary complement to a scientific stance, which exclusively hinges on "hard", quantitative methods in order supposedly to achieve absolute objectivity. The implication of hermeneutics for model validation is that it recognizes the pertinence of subjective judgment. In this connection, interpretive discourses play a crucial role in group model-building and validation. Such discourses lead beyond the subjective, entailing the creation of inter-subjective, shared realities. We will return to this factor in Sect. "Validation Process".

Critical Rationalism

Critical rationalism is a philosophical position founded by Karl R. Popper [19,20]. It grew out of positivism but rejected its verificationist stance. Critical rationalism posits that, in the social domain, theories can never be definitely proved, but can only reach greater or lesser levels of truth. Scientific proofs are confined to the realm of the formal sciences, namely logic and mathematics.

As Popper demonstrates, all theories are provisional. As a consequence, the main criterion for the assessment of a theory's truth status is *falsification* [19]. A theory holds as long as it is not refuted. Consequently, any theory can be upheld as long as it passes the test of falsification. In other words, the fertile approaches to science are not those of corroboration, but the falsificationist efforts to test if theories can be upheld. In the context of modeling this means that validation must undertake attempts to falsify a model, thereby testing its robustness.

Even Popper's theory of science is not unchallenged. For example, Kuhn has made the point that its principles are applicable only to normal science, which operates incrementally within a given paradigm, but not to anomalous science, which uncovers unsuspected phenomena in periods of scientific revolution [12]. This observation has an implication for model validation: Alternative and even multiple model designs should be assessed for their ability to account for fundamental change.

On the Meaning of Validity and Validation

One of the predominant convictions about science is the obsessive idea that proofs are the touchstone of the validity of both theories and models. We follow a different rationale, reverting to the philosophy of science as embodied in critical rationalism.

Popper's refutationist concept (as opposed to a verificationist concept) of theory-testing implies both an evolutionist perspective and an empiricist stance. The evolutionist perspective is primary because it welcomes the challenges posed to a theory, since these attempts at falsification lead to an evolutionary process: successful falsification efforts result in revisions and improvements of the theory. Correspondingly, empiricism is paramount in the social sciences, because the main source for the refutation of a theory is empirical evidence. However, falsification can also be grounded in logical arguments where empirical evidence cannot be obtained. In this sense, a structuralist approach as used in System Dynamics validation transcends the bounds of logical empiricism.

As a consequence of the evolutionist perspective, there is no such thing as absolute validity. Validity is always imperfect, but it can be improved over time. The empiricist aspect of theory-building implies that theories must be validated by means of empirical data. However, logical assay, estimation and judgment are complementary to this empiricist component (see below).

A validation process is about gradually building confidence in the model under study [2]. This is both analytical and synthetic. It is directed at the model as a whole as much as it is at the components of the model. The touchstone of validity is less whether the model is right or wrong: as Sterman states, "… all models are wrong." [28]. Some models, however, fulfill the purpose ascribed to them, i. e., they are useful. Models are inherently incomplete; they cannot claim to be true in an absolute sense, but only to be relatively true [4]. In this sense, *validation* is a *goal-oriented* activity and *validity* a *relative* concept.

Finally, the validation process often involves several people because the necessary knowledge is distributed. In these cases, the dialectics of propositions and refutations, as well as the interaction of different subjective viewpoints, and consensus-building, are integral. Validation processes, then, are semiformal, discursive social procedures with a holistic as opposed to a fragmentary orientation [ibidem].

On Objectivity

If subjective views and judgments are as prominent as alleged above, does objectivity play a role at all? Operational philosophy shows a way out of this dilemma: Rapoport defines objectivity as "invariance with respect to different observers." [21]. Popper has a similar stance in proposing that general statements must be formulated in a way that they can be criticized and, where applicable, falsified [20]. This concept of objectivity is a challenge to model validation: When defining concepts and functions, one must first of all strive for falsifiable statements. In principle, formal models meet this criterion: each variable and every function or relationship can be challenged. And they must be challenged, so that their robustness can be tested. The duty, then, is in finding the invariances that are inter-subjectively accepted as the best approximations to truth. Frequently this is best achieved in group model-building processes [30]. Finally, truth is something we search for but do not possess [20], i. e., even an accepted model cannot guarantee truth with final certainty.

Validation Methods

A considerable set of qualitative and quantitative tests has been developed for the enhancement of model validity. The state-of-the-art has been documented in seminal publications [2,4,7,8,14,17,28]. Our purpose here is to present and exemplify the different tests to encourage and help those who strive to develop high-quality System Dynamics models.

In the following, an overview of the types of tests developed for System Dynamics models is given, without any claim to completeness. Most of these tests have been documented extensively in [2,7,8,28]. The descriptions of the tests adhere closely to the specifications of these authors (mainly Forrester and Senge). In addition, we have developed a new category for tests that concentrate on the context in which the model is to be developed. High-quality models can be created only if the relevant context is taken into consideration. To facilitate orientation, we have attached an overview of all described tests in the Appendix.

In this section we describe three groups of tests: those related to model-related context, tests of model structure and tests of model behavior. Many of the tests described in the following can be utilized for explanatory analysis which aims at an understanding of the problematic behavior of the issue under study. Others are suitable for normative ends, in analyzes targeted on improvements of system performance with regard to a specified objective of the reference system. Also known as policy tests, or policy analyses, these "tests of policy implications differ from other tests in their explicit focus on comparing changes in a model and in the corresponding reality. Policy … tests attempt to verify that response of a real system to a policy change would correspond to the response predicted by the model" [8]. Policy testing can show the risk involved in adopting the model for policy making.

Tests About the Model-Related Context

These tests deal with aspects related to the situation in which the model is to be developed and embedded. They imply metalevel decisions which have to be taken in the first place, before engaging in model-building. Applied ex-post-facto, i. e., after modeling, they allow for assessing the utility of the modeling endeavor as such.

Issue Identification Test. The *raison d'être* of a System Dynamics model is its ability to adequately address an issue and to enhance stakeholders' understanding, an ability which may lead to policy insights and system improvements. The issue identification test examines whether or not the identified issue or problem is indeed meaningful. Has the "right" problem been identified? Does the problem statement address the origins of an issue or only superficial *symptoms*? Whenever complex issues are addressed by a model, different perspectives (e. g. professional, economic, political) must be integrated for accurate problem identification and modeling. This is not a "one-shot-only" test; it must be applied recurrently during the modeling procedure. By reflecting regularly on the correctness of the identified issue, the modeler can increase the likelihood of capturing the origins of suboptimal system behavior.

Adequacy of Methodology Test. Simulation models respond to the limitations of humans' mental ability to comprehend complex, dynamic feedback systems [27]. The adequacy of methodology test scrutinizes whether the System Dynamics methodology is best-suited for dealing with the issue under study. One needs to clearly ascertain if that issue is characterized by dynamic complexity, feedback mechanisms, nonlinear interdependency of structural elements and delays between causes and effects. One needs to ask also if the issue under study could be better addressed by another methodology. For example, in a case where the question is to understand the difference in numerical outcomes between two configurations of a production system, it lets one determine whether discrete event simulation would fulfill this requirement more accurately than System Dynamics.

System Configuration Test. This test asks the fundamental question about whether the structural configuration chosen can be accepted. It challenges the assumption that the model represents the actual working of the system under study. The applicability of a different design would be suggested by its ability to capture new conditions, such as different system configurations, phenomena or rules of the game. Even revolutionary changes might be considered. Such an outlook may require a totally new model, or an alternative model designed from a different vantage point. This would at least feasibly approximate the need to take paradigmatic change into account.

System Improvement Test. The purpose of modeling is to understand a part of reality and to resolve an issue.

The system improvement test can be performed only after the modeling project (an ex-post-facto test), once the insights derived from the model have already been implemented in the real system. This test reestablishes the connection between the abstract mathematical model and the real system. The system improvement test helps to evaluate whether or not model development was successful. In operational terms, any improvements of the real system under study must be compared with explicit objectives. In practice, the test might assess the impact of the modeling process or the model use either on the mental models of decision makers or on changes in organization structures. In principle, assessing the impact of a modeling endeavor is very difficult (one preliminary example is provided by Snabe and Grössler [25]).

Tests of Model Structure

Tests of model structure refer to the "nuts and bolts" of System Dynamics modeling, i. e., to the formal concepts and interrelationships which represent the real system. Model structure tests aim to increase confidence in the structure of the created theory about the behavior mode of interest. The model structure can be assessed by means of either direct or indirect inspection. Tests of model structure assess whether the logic of the model is attuned to the corresponding structure in the real world. They do not yet compare the model behavior with time series data from the real system.

Direct Structure Tests Direct structure tests assess whether or not the model structure conforms to relevant descriptive knowledge about the real system or class of systems under study. By means of direct comparison, they qualitatively assess any disparities between the original system structure and the model structure.

Structure Examination Test. Examination in this case means comparison in the sense just outlined. Qualitative or quantitative information about the real system structure can be obtained either empirically or theoretically. Empirically based tests include reviews of model assumptions about system elements and their interdependencies, e. g., reviews made by highly knowledgeable experts of the real system. Theory-based tests compare the model structure with theoretical knowledge from literature about the type of system being studied. Thereby, a preference for theoretical knowledge specific to the modeled situation over more abstract and general knowledge is usually the case.

To pass the structure examination test, a model must not contradict either the evidence or knowledge about the structure of the real system. This test ensures that the model contains only those structural elements and interconnections that are most likely extant in the real system. In this context, formal inspections of the model's equations, reviews of the syntax for the stock and flow diagram, and walkthroughs along the causal loop diagrams and their embodied causal explanations may be indicated. The experienced reader might recommend the use of statistical tests to identify and validate model structure. As Forrester and Senge [8] indicate, a long-standing discussion exists about the application of inferential statistical tests for structure examination. After a series of experiments, Forrester and Senge conclude "that conventional statistical tests of model structure are not sufficient grounds for rejecting the causal hypotheses in a system dynamics model." [8]. In the future, however, new statistical approaches might enrich the testing procedures.

Parameter Examination Test. A parameter is a quantity that characterizes a system and is held constant in a case under study, but may be varied in different cases (e. g., energy consumption per capita per day). The aim of parameter examination is to evaluate a model's parameters against evidence or knowledge about the real system. The test can utilize both empirical and theoretical information. Furthermore, the test can be conceptual or numerical. The conceptual parameter examination test is about construct validity; it identifies elements in the real system that correspond to the parameters of the model. Conceptual correspondence means that the parameters match elements of the real system's structure. Numerical parameter examination checks to see if the quantities of the conceptually confirmed parameters are estimated accurately. Techniques for the estimation of parameters are described in [9].

Direct Extreme Condition Test. Extreme conditions do not often occur in reality; they are exceptions. The validity of a model's equations under extreme conditions is evaluated by assessing the plausibility of the results generated by the model equations against the knowledge about what would happen under a similar condition in reality. Direct extreme condition testing is a mental process and does not involve computer simulation. Ideally, it is applied to each equation separately. It consists of assigning extreme values to the input variables of each equation. The values of the output variables are then interpreted in terms of what would happen in the real system under these extreme conditions. For example, if a population is zero, then neither births, deaths, nor consumption of resources can occur.

Boundary Adequacy Structure Test. Boundary adequacy is given if the model contains the relevant structural relationships that are necessary and sufficient to satisfy a model's purpose. Consequently, the boundary adequacy test inquires whether the chosen level of aggregation is appropriate and if the model includes all relevant aspects of structure. It should ensure that the model contains the concepts that are important for addressing the problem endogenously. For instance, if parameters are likely to change over time, they should be endogenized [8]. The pertinent validation question is: "Should this parameter be endogenized or not?" That question must be decided in view of the model's purpose.

The boundary adequacy test can be applied in three ways: as a structural test, as a behavioral test, and as a policy test. The names are correspondingly: boundary adequacy structure test, boundary adequacy behavior test, and boundary adequacy policy test.

As a test of model structure, the boundary adequacy test involves developing a convincing hypothesis relating the proposed model structure to the particular issue addressed by the model. The boundary adequacy behavior/policy test (explained in Subsect. "Indirect Structure Tests") continues this line of thinking.

Dimensional Consistency Test. This test checks the dimensional consistency of measurement units of the expressions on both sides of an equation. The test is performed only at the equation level. When all tests of the individual equations are passed, a large system of dimensionally consistent equations results. This test is passed only if consistency is achieved without the use of parameters that have no meaning in respect to the real world. The dimensional consistency test is a powerful test to establish the internal validity of a model.

Indirect Structure Tests Indirect structure tests assess the validity of the model structure indirectly by examining model-generated outcome behaviors. These tests require computer simulation. The comparative activities in these tests are based on logical plausibility considerations which in turn are based on the mental models of the analyst. Comparisons of model generated data and time series about the real system are not yet involved. The tests can be applied to different degrees of model completeness, i. e., to the smallest "atomic" model components, to sub-models, as well as to the entire model.

Indirect Extreme Condition Test. For this test, the modeler assigns extreme values to selected model parameters and compares the generated model behavior to the observed or expected behavior of the real system under the same extreme conditions. This test is the logical continuation of the direct extreme condition test, i. e., many of the extreme conditions mentally developed in the previous stage can now be deployed to evaluate the simulated behavioral consequences. This test can be used for the explanatory analysis phase of modeling, but also for the normative phase of policy development. In the first instance, indirect extreme conditions are used to develop a structure that can reproduce the system behavior of interest and guard against developments impossible in reality. In the latter instance, the introduction of policies aims to improve the system's performance. The indirect extreme policy test introduces extreme policies to the model and compares the simulated consequences to what would be the most likely outcome of the real system if the same extreme policies would have been implemented.

Behavior Sensitivity Test. Sensitivity analysis assesses changes of model outcome behavior given a systematic variation of input parameters. This test reveals those parameters to which the model behavior is highly sensitive, and asks if the real system would exhibit a similar sensitivity to changes in the corresponding parameters. "The behavior sensitivity test examines whether or not plausible shifts in model parameters can cause a model to fail behavior tests previously passed. To the extent that such alternative parameter values are not found, confidence in the model is enhanced." [8]. A model can be numerically sensitive, i. e., the numerical values of variables change significantly, but the behavioral patterns are conserved. It can also exhibit behavioral sensitivity, i. e., the modes of model behavior change remarkably based on systematic parameter variations (Barlas [3] defines several distinct patterns of model behavior).

As the test for indirect extreme conditions, the behavior sensitivity test can also be deployed to assess policy sensitivity. It can reveal the degree of robustness of model behavior and hence indicate to what degree model-based policy recommendations might be influenced by uncertainty in parameter values. If the same policies would be recommended regardless of parameter changes over a plausible range, risk in using the model would be lower than if two plausible sets of parameters lead to distinct policy recommendations.

Integration Error Test. Integration error is the deviation between the analytical solution of differential equations and the numerical solution of difference equations. This test ascertains whether the model behavior is sensitive to changes in either the applied integration method or the chosen integration interval (often referred to as simulation time step). Euler's method is the simplest numerical technique for solving ordinary differential and difference equations. For models that require more precise integration processes, the more elaborated Runge–Kutta integration methods can produce more accurate results, but they require more computational resources.

Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test. The logic for testing boundary adequacy has already been developed under the aspect of direct structure testing in the preceding section. The indirect structure version of this test asks whether model behavior would change significantly if the boundary were extended or reduced; i. e., the test involves conceptualizing additional structure or canceling unnecessary structure with regard to the purpose of the study. As one example of expanding the model boundary, this version of the test allows one to detail the treatment of model assumptions considered as unrealistically simple but still important for the model's purpose. On the other hand, simplifying the model is also a way to reduce the model boundary. The loop-knockout analysis is a useful method to implement this two-sided test. Knockout analysis checks behavior changes induced by the connection and disconnection of a portion of the model structure, and helps the modeler to evaluate the usefulness of those changes with respect to the model's purpose.

The other version of this test is the boundary adequacy policy test. It examines whether policy recommendations would change significantly if the boundary were extended (or restricted): That is, what would happen if the boundary assumptions were relaxed (or confined)?

Loop Dominance Test. Loop dominance analysis studies the internal mechanisms of a dynamic model and their temporal, relative contribution to the outcome behavior of the model. The relative contribution of a mechanism is a complex quantitative statement that explains the fraction of the analyzed behavior mode caused by the mechanism considered in ► System Dynamics, Analytical Methods for Structural Dominance Analysis in. The analysis reveals the relative strengths of the feedback loops in the model. The loop dominance test compares these results with the modeler's or client's assumption about which are the dominant feedback loops in the real system. Since the results are analytical statements, interpretation and comparison with the real system requires profound knowledge about the system under study.

Loop dominance analysis reveals insights about a model on a different level of analysis than the other validation tests discussed so far: It works not on the level of individual concepts or behaviors of variables but on the level of causal structure, and compares the temporal significance of the different structures to each other. The use of this test for model validation is a novelty. If the relative loop dominances of the model map the relative loop dominances of the real system, confidence in the model is enhanced. If the relative loop dominances of the real system are not known, it is still possible to evaluate whether or not the loop dominance logic in the model is reasonable.

Tests of Model Behavior

Tests of model behavior are empirical and compare simulation outcomes with data from the real system under study. On that basis, inferences about the adequacy of the model can be made. The empirical data can either be historical or refer to reasonable expectations about possible future developments.

Behavior Reproduction Tests The family of behavior reproduction tests examines how well model-generated behavior matches the observed historical behavior of the real system. As a principle, models should be tested against data not only from periods of stability but also from unstable phases. Policies should not be designed or tested on the premise of normality, but rather should be validated with a view toward robustness and adaptiveness.

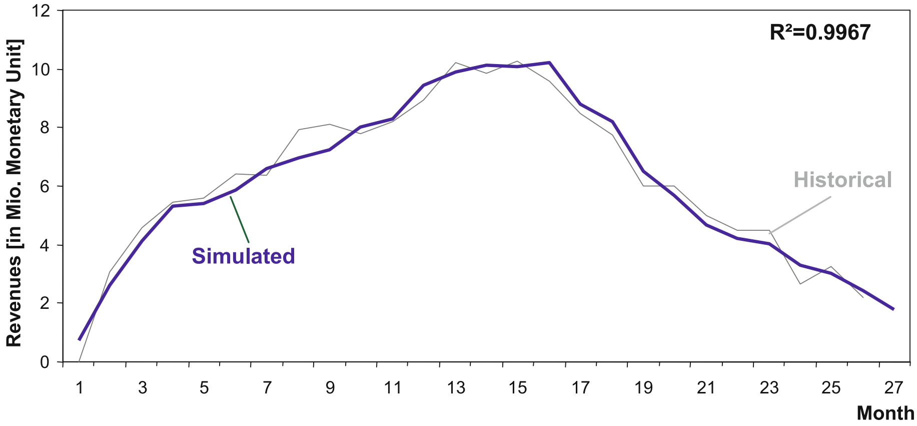
Symptom Generation Test. This test indicates whether or not a model produces the symptom of difficulty that motivated the construction of the model. To pass the symptom generation test is a prerequisite for considering policy changes, because "unless one can show how internal policies and structures cause the symptoms, one is in a poor position to alter those causes" [8].

Summary statistics, which measure and enable the interpretation of quantitative deviations, provide the means to operationalize the symptom generation test.

One known example is Theil inequality statistics, which measures the mean square-error (MSE) between the model-generated behavior and the historical time series data. It breaks down the deviation into three sources of error: Bias (U_m), unequal variation (U_s), and unequal covariation (U_c) [26].

An example taken from Schwaninger and Groesser [22] illustrates the interpretation of the error sources.

This example from an industrial firm concerns the design of a model that replicates the observed, historical product life-cycle pattern with high accuracy (Fig. 2). "Product Revenue" is the main variable of interest and specifies the symptom (growth phase followed by rapid decay). The mean square-error for revenues is 0.35. The



System Dynamics Modeling: Validation for Quality Assurance, Figure 2
An example comparison of historical and simulated time series for product revenues. The explained variance is close to 100% (R^2 = 0.9967)

individual components of the inequality statistics are: $U_m = 0.01, U_s = 0.01, U_c = 0.98$. The break down of the statistics shows that the major part of the error is in the U_c component, while the other two sourees of error are small. This signifies that the point-by-point values of the simulated and historical data do not match, even though the model captures the dominant trend and the average values in the historical data. Such a situation indicates that a major part of the error is probably unsystematic, and therefore the model should not be rejected for failing to match the noise component of the data. The residuals of the historic and simulated time series show no significant trend. This strengthens the assessment that the model comprises a structure that captures the fundamental dynamics of the issue under consideration.

Frequency Generation and Phase Relationship Tests. These tests focus on the frequencies of time series and phase relationships between variables. An example is the pattern of investment cycles in an industry. These tests are superior to point-by-point comparisons between model-generated and observed behavior (cf. [7]).

Frequency refers to periodicities of fluctuation in a time series. Phase relationship is the relationship between the time series of at least two variables. In principle, three phase relations are possible: Preceding, simultaneous, and successive. The frequency generation test evaluates whether or not the periodicity of a variable is in accordance with the real system. The phase relationship test assesses the phase shifts of at least two variables by comparing their trajectories.

If the phase shift between the selected simulation variables contradicts the phase shift between the same variables as observed or expected in the real system, a structural flaw in the model might be diagnosed. The test can uncover failures in the model, but offers only little guidance as to where the erroneous part of the model might be. The autocorrelation function test is one way to operationalize the frequency generation test [1]. The function test consists in comparing the autocorrelation functions of the observed and the model-generated behavior outputs, and can detect if significant errors between them exist.

Modified Behavior Test. Modified behavior can arise from a modified model structure or changes in parameter values. This test concerns changes in the model structure. It can be performed if data about the behavior of a structurally modified version of the real system are available. "The model passes this test if it can generate similar modified behavior, when simulated with structural modifications that reflect the structure of the "modified" real system" [2]. The applicability of this test is rather limited since it requires specific data about the modified real system which must be similar in kind to the original real system. Only under this condition can additional insights into the suitability of the original model structure be obtained. If the modified real system deviates strongly from the original real system, the test does not result in any additional insights, because no stringent conclusions about the validity of the original system can be derived from a model that is dissimilar in its structure.

Multiple Modes Test. A mode is a pattern of observed behavior. The multiple mode test considers whether a model is able to generate more than one mode of observed behavior, for instance, if a model about the production sector of an economy generates distinct patterns of fluctuations for the short-term (production, employment, inventories, and prices) and for the long term (investment, capital stock) [15]. " A model able to generate two distinct periodicities of fluctuation observed in a real system provides the possibility for studying possible interaction of the modes and how policies differentially affect each mode" [8].

Behavior Characteristic Test. Characteristics of a behavior are features of historical data that are clearly distinguishable, e. g., the peculiar shape of an oscillating time series, sharp peaks, long troughs, or such unusual events as an oil crisis. Since System Dynamics modeling is not about point prediction, the behavior characteristic test evaluates whether or not the model can generate the circumstances and behavior leading to the event. The creation of the exact time of the behavior is not part of the test.

Behavior Anticipation Tests System Dynamics models do not strive to forecast future states of system variables. Nevertheless, given that the fundamental system structure is not subject to rapid and fundamental change, dynamic models might provide insights about the possible range of future behaviors. Hence, behavior anticipation tests are similar to behavior reproduction tests but possess a higher level of uncertainty.

Pattern Anticipation Test. This test examines whether a model generates patterns of future behavior which are assumed to be qualitatively correct. The limits of anticipation reside in the fact that that the structure of the system may change over time. The pattern anticipation test entails evaluation of periods, phase relationships, shape, or other characteristics of behavior anticipated by the model. One possibility for implementing this test is to split the historical time series into two data sets and introduce an artificial present time at the end of the first data series. The first set is then used for model development and calibration. The second data series is employed to perform the behavior anticipation test, i. e., to evaluate whether the model is able to anticipate possible future behavior.

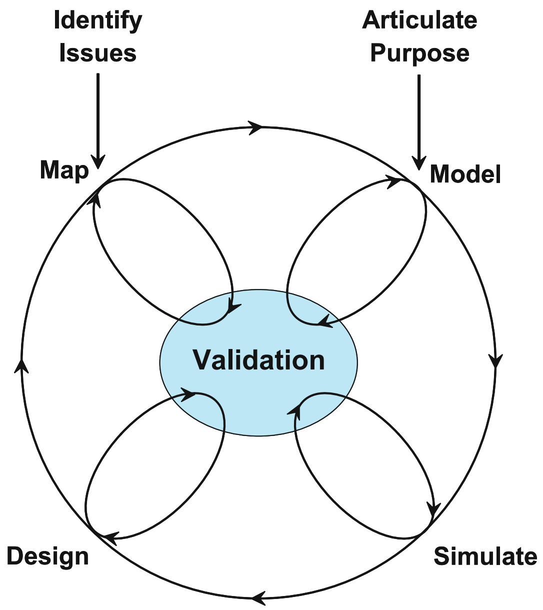
This test can also be used for policy considerations, in which case it is called "Changed Behavior Anticipation Test". It determines whether the model correctly anticipates how the behavior of the real system will change if a governing policy is altered.

Event Anticipation Test. In respect to System Dynamics, the anticipation of events does not imply knowing the exact time at which the events occur; it rather means understanding the dynamic nature of events and being able to identify the antecedents leading to them. For instance, the event anticipation test is passed if a model has the ability to anticipate a steep peak in food prices based on the development of the conditioning factors.

Behavior Anomaly Test In constructing and analyzing a System Dynamics model, one strives to make it behave like the real system under study. However, the analyst may detect anomalous features of the model's behavior which conflict with the behavior of the real system. Once the behavioral anomaly is traced to components of the model structure responsible for the anomaly, one often finds flaws in model assumptions. The test for recognizing behavioral anomalies is sporadically applied throughout the modeling process.

Family Member Test A System Dynamics model often represents a family of social systems. Whenever possible, a model should be a general representation of the class of that system to which the particular case belongs. One should ask if the model can generate the behavior in other instances of the same class. "The family-member test permits a repeat of the other tests of the model in the context of different special cases that fall within the general theory covered by the model. The general theory is embodied in the structure of the model. The special cases are embodied in the parameters. To perform this test, one uses the particular member of the general family for picking parameter values. Then one examines the newly parametrized model in terms of the various model tests to see if the model has withstood transplantation to the special case" [8]. The model should be calibrated so as to be applicable to the widest range of related systems. For the family member test, only the parameter values of the model are subject to alterations; changes in the model structure are part of the modified behavior test, as discussed in the preceding section.

Surprise Behavior Test A surprising model behavior is a behavior that is not expected by the analysts. When such an unexpected behavior appears, the model analysts must first understand the causes of the unexpected behavior within the model. They then compare the behavior and its causes with those of the real system. In many cases, the surprising behavior turns out to be due to a formulation flaw in the model. However, if this procedure leads to the identification of behavior previously unrecognized in the real system, the confidence in the model's usefulness is strongly enhanced. Such a situation may signify a model-based identification of a counter-intuitive behavior in a social system.



System Dynamics Modeling: Validation for Quality Assurance, Figure 3
Validation in the context of the System Dynamics modeling procedure

Turing Test The Turing test is a qualitative test which uses the intuitive knowledge of system experts to evaluate model behavior. Experts are presented with a shuffled collection of real and simulated output behavior patterns. They are asked if they can distinguish between these two types of patterns. If they are unable to discern which pattern belongs to the real system and which to the simulation output, the Turing test is passed. Similar to the phase relationship test, the Turing test is powerful in its ability to indicate structural flaws, but offers only little guidance for locating them in the model.

Validation Process

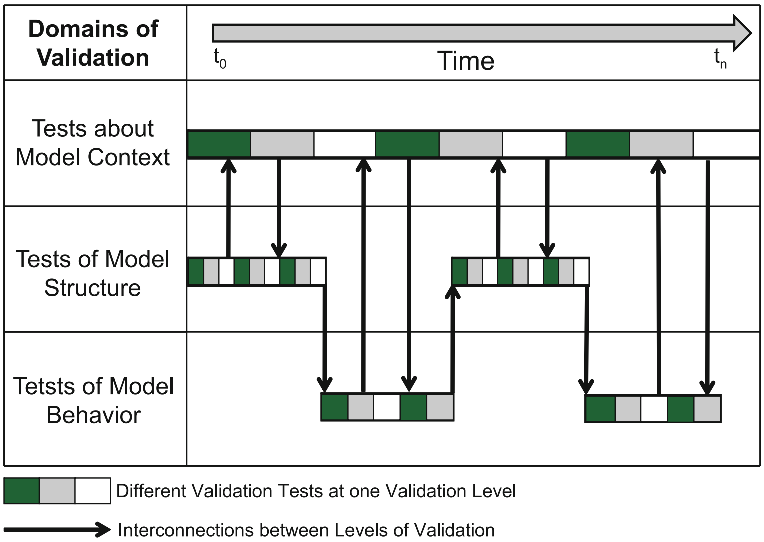
The validation process pervades all phases of model-building and reaches even beyond, into the phases of model implementation and use. The diagram in Fig. 3 visualizes the function of validation in the process of model-building.

For the purposes of this contribution, validation is placed at the center of the scheme. From there it is dispersed through all steps of the modeling process, Map (high-level model creation), Model (build the formal model), Simulate (explore scenarios, etc.) and Design (articulation of policies). We have limited the differentiation of these steps in order to highlight the structure of the process – a recursive structure drawn as a nested loop line. After the initial identification of issues and the articulation of model purpose, the simplified diagram denotes the four phases, of mapping to modeling to simulation and design. The small loops symbolize micro-processes in which, for example, a model is submitted to validation, e. g., a direct structure test, which may lead to its modification (two small arrows). The larger loops illustrate more comprehensive processes. For example, an indirect structure test of the model is carried out, in which the behavior is tested by means of simulation. Or a policy test by simulation leads to implications for design (large loop), and the design is validated in detail thereafter (small loop).

Now, we should note that the process scheme reminds us of a further aspect which is quite fundamental. If the results of the model's operation, e. g., a "prediction", diverge from the results of a test, then either the model is wrong or the test is inadequate (see p. 168 in [24]). This meta-perspective lets us keep an eye on the adequacy of the tests: is the logic of the test flawless? Are the data sources in order? (see adequacy of methodology test in Sect. "Validation Methods").

Model-building is a process of knowledge-creation, and model validation is an integral part of it. As the model is validated using the methods described in the former chapter, insights emerge, and a better understanding of the system under study keeps growing. But model-building is also a construction of a reality in the minds of observers [31,32] concerned with an issue. In this procedure, validation is supposed to be a "guarantor" for the realism of the model, a control function for preventing gross aberrations in individual and collective perceptions. Validation should encompass precautions against cognitive limitations and modeler blindness. The set of tests presented above is a system of heuristic devices for enhancing such provisions. A question not yet answered is how these tests should be ordered along the timeline. We have fleshed out three structural principles, which are illustrated in Fig. 4:

1. *Validation is a parallel process:* Validation in all three domains – context, structure and behavior – is carried out in a synchronized fashion, as shown in Fig. 4. Context validation is continuous, while the other two components show alternations.
2. *Parts of the validation process have a sequential structure:* This refers to the alternations between the components of structure and behavior validation. In principle, they occur alternately, with structural validation taking the lead and behavior validation following. After that, one might revert to structural validation again and so forth.


Domains of Validation
t0
Time
tn
Tests about Model Context
Tests of Model Structure
Tetsts of Model Behavior
Different Validation Tests at one Validation Level
Interconnections between Levels of Validation

System Dynamics Modeling: Validation for Quality Assurance, Figure 4
The interplay of validation activities

3. *Validation processes are polyrhythmic:* The length and accentuation of validation activities vary among the three levels. This fact is symbolized by the frequency of the vertical lines in the blocks of the chronogram.

A further important factor affecting the validation process is the degree of resolution: micro, meso or macro (as visualized in Fig. 1). The focus of validation is primarily on micro-objects, the smallest building blocks of a model, for example, a stock or a subsystem containing a stock with its flows. One could call them metaphorically *atoms* or *molecules*. Each building-block should be validated individually, before it is integrated into the overall model structure. The reason is that at this atomic level disfunctionalities or errors of thinking are discovered immediately, while at higher levels of resolution the identification of structural flaws is more difficult and cumbersome. The same holds for the relation between modules (meso) and the whole model (macro). Before adding a module, it should be validated in itself. This way, errors at the level of the whole system can be minimized and, it is very important to add, counterintuitive behavior of the model can be understood with more ease.

Until now we have examined what occurs in a validation process and how the process is structured. Finally, we raise the issue of who the actors are and why. In this context, we will concentrate on group processes in model validation.

Different observers associate diverse contents with a system, and they might even conceive the system distinctly, as far as its boundaries, goals and structures are concerned. They might also succumb to erroneous inferences and therefore adhere to defective propositions. Consequently, error-correcting devices are needed. A powerful mechanism for this purpose is the practice of model-building and validation in groups. We have already referred to that concept in respect to several of the methods discussed in Sect. “Validation Methods”, and now we will briefly expand on it.

Group Model-building (GMB) is a methodology to facilitate team learning with the help of System Dynamics [30]. The methodology consists of a set of methods and instruments as well as heuristic principles. These are meant to facilitate the elicitation of knowledge, the negotiation of meanings, the creation of a shared understanding of a problem in a team, as well as the joint construction and validation of models. The process of GMB is essentially a dialog in which different interpretations of the real system under study are exposed, transformed, aligned and translated into the concepts and relationships which make up the model system. This is mainly a matter of structural validation, of qualitative mapping and the elaboration of the formal model.

Given its transdisciplinary approach, GMB enables an integration of different perspectives into one shared image of the system-in-focus. GMB is an important provision for attaining higher model quality: it can broaden the available knowledge base, inhibit errors and show itself to be a cohesive force in the quest for consensual model validation. The opportunity for validation inheres in the broad knowledge base normally available in a modeling group. Much of this knowledge can be leveraged for validation purposes. Most validation tests are carried out in coordination with model-building activities. Often the tests become a task to be accomplished between workshops. However, the members of the model-building group can, in principle, be made available for knowledge input into and monitoring of validation activities.

A functioning GMB process requires a number of necessary elements [18]: commitment of key players (e. g., attendance of workshops), impartial facilitation, on-the-spot modeling at conversational pace, with continuous display of the developing model as well as an interactive and iterative group process.

Let us not forget that there are many situations in which one single person is in charge of building and validating a model. In these cases the modeler must constantly challenge his or her own position. Normally, it is preferred that one should also call for external judgment in reviews, walkthroughs and the like. The same holds for knowledge supply. One-person modelers can find a lot of material in the media, libraries, the internet, etc., but it is also usually beneficial to find experienced persons from whom to elicit relevant knowledge, or even persons who join the modeling and validation venture.

Synopsis and Outlook

Models should be relevant for coping with the complexity of the real world. At the same time, the methods by which they are constructed must be rigorous; otherwise the quality of the model suffers. Rigor and relevance are not entirely dichotomous, but given resource constraints they are in competition to a certain extent. Lack of rigor in building a model is often worse than limitations to the model's relevance. One may say, *cum grano salis*: incomplete validation entails complete irrelevance. Modelers must find a way to ensure both rigor and relevance, as both are necessary conditions for achieving the model purpose. Neither alone is sufficient, but one may assume that, taken together, rigor and relevance are sufficient conditions. The relative importance of these two dimensions of model building may vary over time as a function of the model quality achieved. At the beginning, relevance might be more important, while at high levels of model accomplishment rigor might become prevalent.

Investing in high model quality is indeed both worthwhile and imperative. It is impressive to register the fact that model validation has achieved higher levels of rigor not only in the academic field but also in the world of affairs: According to Coyle and Exelby, the need for orientating decisions about "real-world" affairs has also fueled strong efforts among commercial modelers and consultants for ensuring model validity [5].

We have discussed two essential aspects of model validation, the epistemological foundations and methodological procedures for ensuring model validity. The main conclusion we have reached on epistemology is that crude positivism has been superseded by newer philosophical orientations that provide guidance for an adequate concept of validation in System Dynamics. Validation has been defined as a rich and well-defined process by which the confidence in a model is gradually enhanced. Validity, then, is always a matter of degree, never an absolute property.

Well-defined here is not meant in the sense of a rigid algorithm, but as the rigorous application of a battery of validation methods which we have described in some detail. We have included a number of new validation tests by which modelers' understanding of the relevant context can be scrutinized. These additional tests are rightly supposed to prevent wrong methodological choices. They should also trigger innovative approaches to the issues under study and foster the ability to think in terms of contingencies. Finally, they should liberate modelers from tunnel vision and open avenues to creativity. The imperative here is to cultivate a "sense of the possible" (Robert Musil's *Möglichkeitssinn*) and a skepticism against the supposedly impossible (see also [29]).

Simulation based on formal dynamic models is likely to become ever more important for both private and public organizations. It will continue to support managers at all levels in decision-making and policy design. The more that models are relied upon, the greater the importance of their high quality. Therefore, model validation is one of the big issues lying ahead in System Dynamics modeling.

Appendix: Overview of the Tests Described in This Chapter

1. **Tests of the Model-Related Context**
   - 1.1 Issue Identification Test
   - 1.2 Adequacy of Methodology Test
   - 1.3 System Configuration Test
   - 1.4 System Improvement Test
2. **Tests of Model Structure**
   - 2.1 Direct Structure Tests
     - 2.1.1 Structure Examination Test
     - 2.1.2 Parameter Examination Test
     - 2.1.3 Direct Extreme Condition Test

2.1.4 Boundary Adequacy Structure Test
2.1.5 Dimensional Consistency Test
2.2 Indirect Structure Tests
2.2.1 Indirect Extreme Condition Test
2.2.2 Behavior Sensitivity Test
2.2.3 Integration Error Test
2.2.4 Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test
2.2.5 Loop Dominance Test

3. **Tests of Model Behavior**

3.1 Behavior Reproduction Tests
3.1.1 Symptom Generation Test
3.1.2 Frequency Generation and Phase Relationship Test
3.1.3 Modified Behavior Test
3.1.4 Multiple Modes Test
3.1.5 Behavior Characteristic Test
3.2 Behavior Anticipation Tests
3.2.1 Pattern Anticipation Test
3.2.2 Event Anticipation Test
3.3 Behavior Anomaly Test
3.4 Family Member Test
3.5 Surprise Behavior Test
3.6 Turing Test

Bibliography

Primary Literature

1. Barlas Y (1990) An autocorrelation function test for output validation. Simulation 55(1):7–16
2. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
3. Barlas Y (2006) Model validity and testing in System Dynamics: Two specific tools. Paper presented at the 24th International Conference of the System Dynamics Society, Nijmegen
4. Barlas Y, Carpenter S (1990) Philosophical roots of model validity – two paradigms. Syst Dyn Rev 6(2):148–166
5. Coyle G, Exelby D (2000) The validation of commercial system dynamics models. Syst Dyn Rev 16(1):27–41
6. Feyerabend P (1993) Against method, 3rd edn. Verso, London
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
8. Forrester JW, Senge PM (1980) Test for building confidence in System Dynamics models. In: Legasto AA Jr, Forrester JW, Lyneis JM (eds) System Dynamics. North-Holland Publishing Company, Amsterdam, pp 209–228
9. Graham AK (1980) Parameter estimation in system dynamics modeling. TIMS Studies in the Management Sciences 14:125–142
10. Heracleous L (2006) Discourse, interpretation, organization. Cambridge University Press, Cambridge
11. James W (1987) Writings 1902–1910. Library of America, New York
12. Kuhn T (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago
13. Lacey AR (1996) A dictionary of philosophy, 3rd revised edn. Barnes and Noble, New York
14. Lane DC (1995) The folding star: A comparative reframing and extension of validity concepts in System Dynamics. In: Simada T, Saeed K (eds) Proceedings of 1995 international System Dynamics conference, 30 July–4 Aug, vol I. System Dynamics Society, Lincoln, pp 111–130
15. Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge
16. Mattheij RMM, Rienstra SW, Boonkkamp JH MtT (2005) Partial differential equations: Modeling, analysis, computation. Society for Industrial and Applied Mathematics (SIAM), Eindhoven
17. Petersen DW, Eberlein RL (1994) Understanding models with vensim. In: Morecroft JDW, Sterman JD (eds) Modeling for learning organiziations. Productivity Press, Portland, pp 339–358
18. Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) Advances in decision analysis. From foundations to applications. Cambridge University Press, Cambridge, pp 375–399
19. Popper KR (1959) The logic of scientific discovery. Basic Books, New York (latest edition: 2002, Routledge, London)
20. Popper KR (1972) Objective knowledge: An evolutionary approach. Clarendon Press, Oxford
21. Rapoport A (1954) Operational philosophy. Integrating knowledge and action. Harper, New York
22. Schwaninger M, Groesser SN (2008) Model-based theory-building with system dynamics. Syst Res Behav Sci 25:1–19
23. Seiffert H, Radnitzky G (1994) Handlexikon der Wissenschaftstheorie, 2nd edn. DTV Wissenschaft, Munich
24. Smith VL (2008) Rationality in economics: Constructivist and ecological forms. Cambridge University Press, Cambridge
25. Snabe B, Grössler A (2006) System dynamics modelling for strategy implementation – case study and issues. Syst Res Behav Sci 23(4):467–481
26. Sterman JD (1984) Appropriate summary statistics for evaluating the historical fit of system dynamics models. Dynamica 10(2):51–66
27. Sterman JD (1989) Misperceptions of Feedback in Dynamic Decision Making. Organ Behav Human Decis Process 43(3):301–335
28. Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston
29. Taleb NN (2007) The black swan. The impact of the highly improbable. Random House, New York
30. Vennix JAM (1996) Group model building: Facilitating team learning using System Dynamics. Wiley, Chichester
31. von Foerster H (1984) Observing systems, 2nd edn. Intersystems Publications, Seaside
32. von Glasersfeld E (1991) Abschied von der Objektivität. In: Watzlawick P, Krieg P (eds) Das Auge des Betrachters. Piper, Munich, pp 17–30

Books and Reviews

Finlay PN (1997) Validity of decision support systems: Towards a validation methodology. Syst Res Behav Sci 14(3):169–182

Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge

Law AM (2007) Simulation modeling and analysis, 4th edn. McGraw-Hill, New York

Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System Dynamics. North-Holland, Amsterdam

Morecroft J (2007) Strategic modelling and business dynamics: A feedback systems approach. Wiley, Chichester

Sargent RG (2004) Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (eds) Proceedings of the 2004 winter simulation conference. ACM-Association for Computing Machinery, Washington DC, pp 17–28

Sterman JD (2001) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston

Warren K (2008) Strategic management dynamics. Wiley, Chichester

System Dynamics Models of Environment, Energy and Climate Change

Andrew Ford
School of Earth and Environmental Sciences,
Washington State University, Pullman, Washington,
USA

Article Outline

Glossary

CO_2 Carbon dioxide is the predominant greenhouse gas. Anthropogenic CO_2 emissions are created largely by the combustion of fossil fuels.

CGCM Coupled general circulation model, a climate model which combines the atmospheric and oceanic systems.

GCM General circulation model, a term commonly used to describe climate models maintained at large research centers.

GHG GHG is a greenhouse gas such as CO_2 and methane. These gases contribute to global warming by capturing some of the outgoing infrared radiation before it leaves the atmosphere.

GT Gigaton, a common measure of carbon storage in the global carbon cycle. A GT is a billion metric tons.

IPCC The Intergovernmental Panel on Climate Change was formed in 1988 by the World Meteorological Organization and the United Nations Environmental Program. It reports research on climate change. Their assessments are closely watched because of the requirement for unanimous approval by all participating delegates.

Definition of the Subject

System dynamics is a methodology for studying and managing complex systems which change over time. The method uses computer modeling to focus our attention on the information feedback loops that give rise to the dynamic behavior. Computer simulation is particularly useful when it helps us understand the impact of time delays and nonlinearities in the system. A variety of modeling methods can aid the manager of complex systems. Coyle (p. 2 in [3]) puts the system dynamics approach in perspective when he describes it as that "branch of control theory which deals with socio-economic systems, and that branch of management science which deals with problems of controllability." The emphasis on controllability can be traced to the early work of Jay Forrester [9] and his background in control engineering [10]. Coyle highlighted controllability again in the following, highly pragmatic definition:

> *System dynamics is a method of analyzing problems in which time is an important factor, and which involve the study of how a system can be defended against, or made to benefit from, the shocks which fall upon it from the out-side world.*

The emphasis on controllability is important as it directs our attention to understanding and managing the system, not to the goal of forecasting the future state of the system. Making point predictions is the objective of some modeling methods, but system dynamics models are used to improve our understanding of the general patterns of dynamic behavior. System dynamics has been widely used in business, public policy and energy and environmental policy making. This article describes applications to energy and environmental systems.

Introduction

System dynamics has been used extensively in the study of environmental and energy systems. This article describes some of these applications, paying particular attention to the problem of global climate change. The applications were selected to illustrate the power of the method in promoting an interdisciplinary understanding of complex problems.

The applications to environmental and energy systems are similar to applications to other systems described in this encyclopedia. They usually begin with the recognition of a dynamic pattern that represents a problem. System dynamics is based on the premise that we can improve our understanding of the dynamic behavior by the construction and testing of computer simulation models. The models are especially helpful when they illuminate the key feedbacks that give rise to the problematic behavior.

System dynamics is explained in the core article in this volume, in the early texts by Forrester [9], Coyle [3] and Richardson [18] and in more recent texts on strategy by Warren [22] and by Morecroft [17]. The most comprehensive explanation is provided in the text on business dynamics by Sterman [19]. Applications to environmental systems are explained in the text by Ford [7]. The most widely read application to the environment is undoubtedly *The Limits to Growth* [16]. Collections of environmental applications appear in special issues of the *System Dynamics Review* [11,20].

The models are normally implemented with visual software such as Stella (http://www.iseesystems.com), Vensim (http://www.vensim.com/) or Powersim (http://www.powersim.com/). These programs use stock and flow icons to help one see where the accumulations of the system take place. They also help one to see the information feedback in the simulated system. The programs use numerical methods to show the dynamic behavior of the simulated system. The examples selected for this article make use of the Stella and Vensim software.

This article begins with textbook examples of environmental resources in the western US. The management of water levels at Mono Lake in Northern California is the first example. It shows a hydrological model to simulate the decline in lake levels due to water exported out of the basin. The second example involves the declining salmon population in the Tucannon River in Eastern Washington. These examples demonstrate the clarity of the approach, and they illustrate the potential for interdisciplinary modeling.

The article then turns to the topic of climate change and global warming. The focus is on the global carbon cycle and the growing concentration of carbon dioxide (CO_2) in the atmosphere. A wide variety of models have been used to improve our understanding of the climate system and the importance of anthropogenic CO_2 emissions. Examples of system dynamics models are presented to show how they can improve our understanding and provide a platform for interdisciplinary analysis.

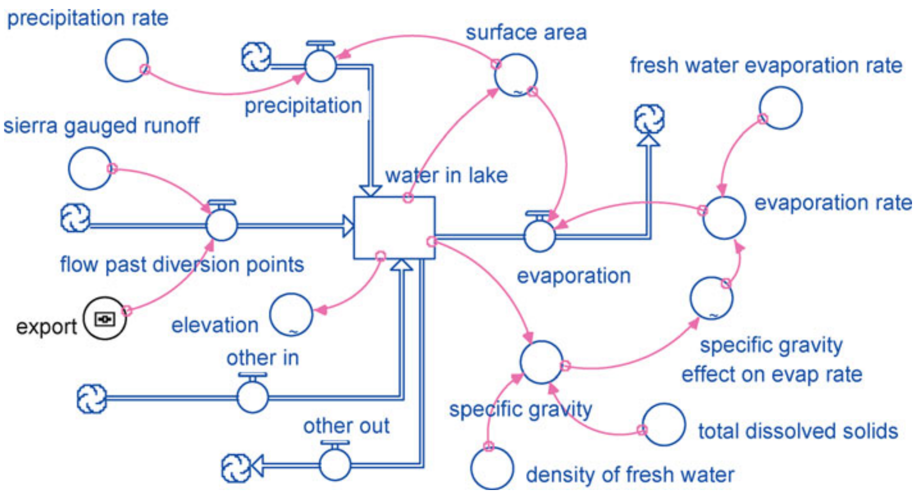
System dynamics has also been widely applied in the study of energy problems, especially problems in the electric power industry. The final section describes two applications to electric power. The first involved the financial problems of regulated electric utilities in the US during the 1970s. It demonstrates the usefulness of the method in promoting an interdisciplinary understanding of the utilities' financial problems. The second study dealt with the CO_2 emissions in the large electricity system in the Western USA and Canada. It demonstrated how the power industry could lead the way in reducing CO_2 emissions in the decades following the implementation of a market in carbon allowances.

The Model of Mono Lake

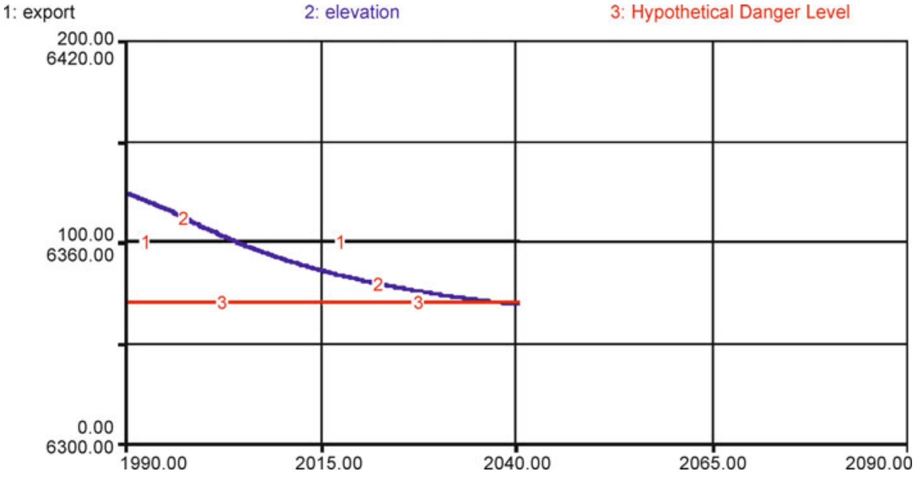
Mono Lake is an ancient inland sea on the east side of the Sierra Nevada Mountains in California. Microscopic algae thrive in its saline waters, and the algae support huge populations of brine flies and brine shrimp which can, under the right conditions, provide a virtually limitless food supply for migratory and nesting birds. Starting in 1941, stream flows toward Mono Lake were diverted into the aqueduct for export to Los Angeles. The large export deprived the lake of the historical flows, and the volume shrunk over the next four decades. By 1980, the lake's volume was cut approximately in half, and its salinity nearly doubled. Higher salinity levels posed risks to the ecosystem, and environmental scientists feared for the future of the lake ecosystem. Various groups filed suit in the 1970s to limit exports, and the California Supreme Court ruled in 1983 that public trust doctrine mandated a reconsideration of the management of the waters of the Mono Basin. That reconsideration led to a long-term plan to limit exports until the lake's elevation would return to safer levels.

Figure 1 shows a system dynamics model to simulate water flows and storage in the Mono Basin. The goal was to understand the pattern of decline over four decades and to study the responsiveness of the lake to a change in export policy. The model is implemented with the Stella software, and Fig. 1 shows how the model appears when using the software. A single stock variable is used to represent the storage in the basin. The main flow into the lake is the flow from gauged streams that bring runoff from the Sierra to the lake. The aqueduct system diverts a portion of this flow south to Los Angeles, and the flow allowed past the diversion points is the main flow into the lake. The main outflow is the evaporation. It depends on the surface area of the lake and the evaporation rate. The surface area depends in a nonlinear way on the volume of water in the lake. Figure 1 shows that this model follows the standard, system dynamics practice of using familiar names to convey the meaning of the variables in the model. (These particular names match the terms used by water managers and hydrological models of the basin.)

Figure 2 shows the simulated decline in the lake if exports were allowed to continue at high levels for 50 years. The lake would decline from 6374 to around 6342 feet above sea level, a value which is designated as a hypothetical danger level for this simulation. The long, gradual



System Dynamics Models of Environment, Energy and Climate Change, Figure 1
Stella diagram of the model of Mono Lake

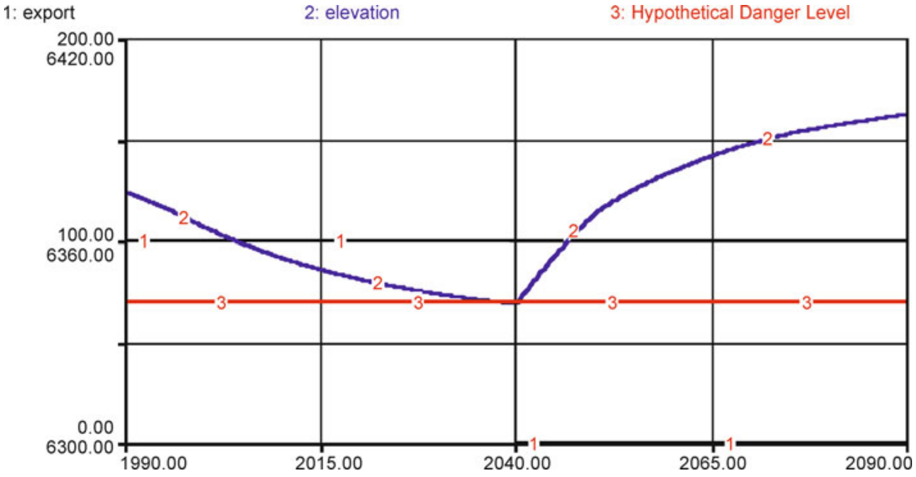


System Dynamics Models of Environment, Energy and Climate Change, Figure 2
Simulated decline in Mono Lake elevation if historical export were allowed to continue until the year 2040

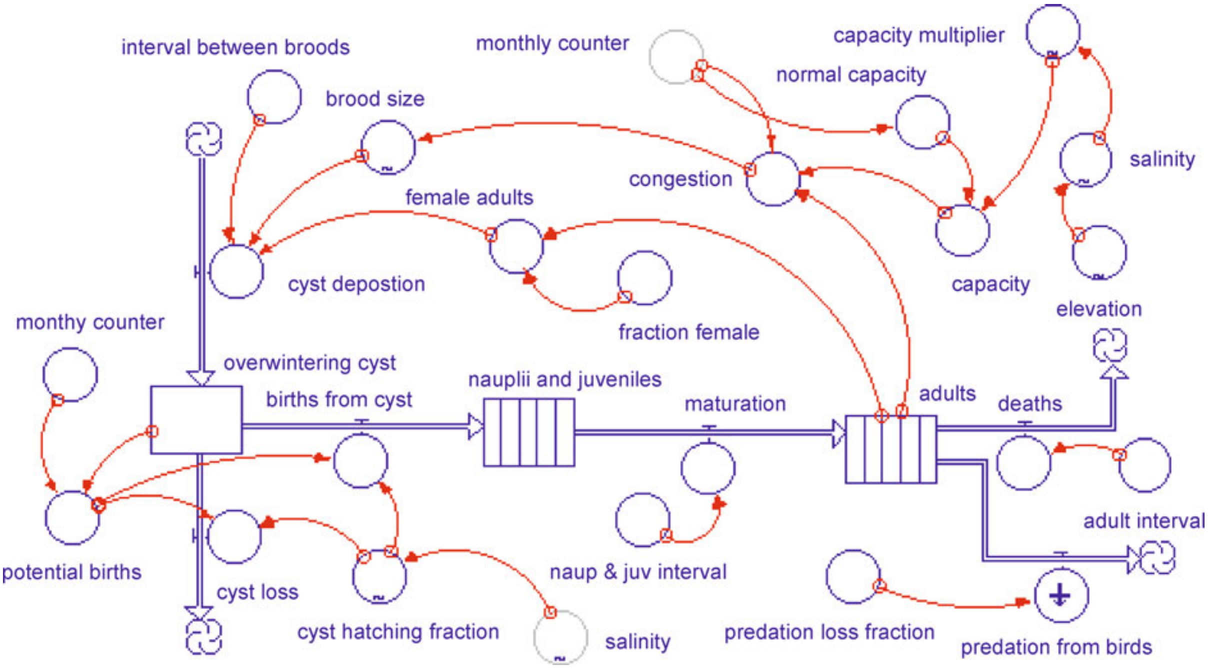
decline is a match of projections by the other hydrological models used in the management plan for the basin. The lake will continue to fall until the area has been reduced sufficiently to create an evaporation which will lead to a balance of the flows in and out of the basin.

Figure 3 shows the simulated responsiveness of the lake to a change in export. The export is cut to zero midway through the simulation, and the elevation increases rapidly in the ensuing decade. The simulation reveals an immediate and rapid response, indicating that there is little downward momentum associated with the hydrology of the basin. This responsiveness is highly relevant to the management plan. When the lake falls to a dangerous level, the export could be reduced, and the lake would climb to higher elevations within a few years after the change in policy. This rapid response supports the "wait and see" argument by those who advocated waiting for full signs of a dangerous salinity before changing export policy.[1] But there is far more than hydrology at work in this system. The waters of Mono Lake support a complex ecosystem which may or may not recover as quickly as the lake elevation. To explore the larger system requires

[1] "Wait and see" may be supported by an analysis of the hydrology of the basin, but it does not necessarily make sense when considering the long delays in the political and managerial process to change water export.



System Dynamics Models of Environment, Energy and Climate Change, Figure 3
Simulated recovery of Mono Lake elevation if export is set to zero for the second half of the simulation



System Dynamics Models of Environment, Energy and Climate Change, Figure 4
Stella model of the brine shrimp population of Mono Lake

an interdisciplinary model, one that looks at both hydrology and population biology.

Figure 4 shows a model of the population of brine shrimp that live in Mono Lake. The life cycle begins when the adult females deposit cysts in the summer. A stock is assigned to the over wintering cysts. The nauplii and juvenile phases are combined into a second stock, and the maturation leads to a new population of adults in the following summer. The model operates in months and is simulated over a long time interval to show the population response to long-term changes in elevation and in salinity. The model shows the population's response to changes in lake elevation, so one can learn about the delays in the population's response to the changes in lake elevation. Since the shrimp life cycle is 12 months, one would expect the population to rebound rapidly after the increase in elevation and the reduction in salinity. The model confirms that the shrimp population would increase rapidly in the years following the elimination of water export from the basin.

The Mono Lake models are textbook models [7]. They demonstrate the clarity that the system dynamics approach brings to the modeling of environmental systems. The stock and flow icons help one see the structure of the system, and the long variable names help one appreciate the individual relationships. The simulation results help one understand the downward momentum in the system. In this particular case, there is no significant downward momentum associated with either the hydrological dynamics or the population dynamics.

The model in Fig. 1 allows for a system dynamics portrayal of the type of calculations commonly performed by hydrologists. Compared to the previous methods in hydrology, system dynamics adds clarity and ease of experimentation. The population model in Fig. 4 is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation in this discipline as well.

The main theme of this article is that system dynamics offers the opportunity for interdisciplinary modeling and exploration. The Mono Lake case illustrates this opportunity with the combination of the hydrological and biological models that allows one to simulate management policies that control export based on the size of the brine shrimp population. The new model is no longer strictly hydrology nor strictly population biology; it is an interdisciplinary combination of both. And by using stock and flow symbols that are easily recognized by experts from many fields of study, the system dynamics enables quick transfer of knowledge. The ability to combine perspectives from different disciplines is one of the most useful aspects of the system dynamics approach to environmental and energy systems. This point is illustrated further with each of the remaining examples in the article.

The Model of the Salmon in the Tucannon River

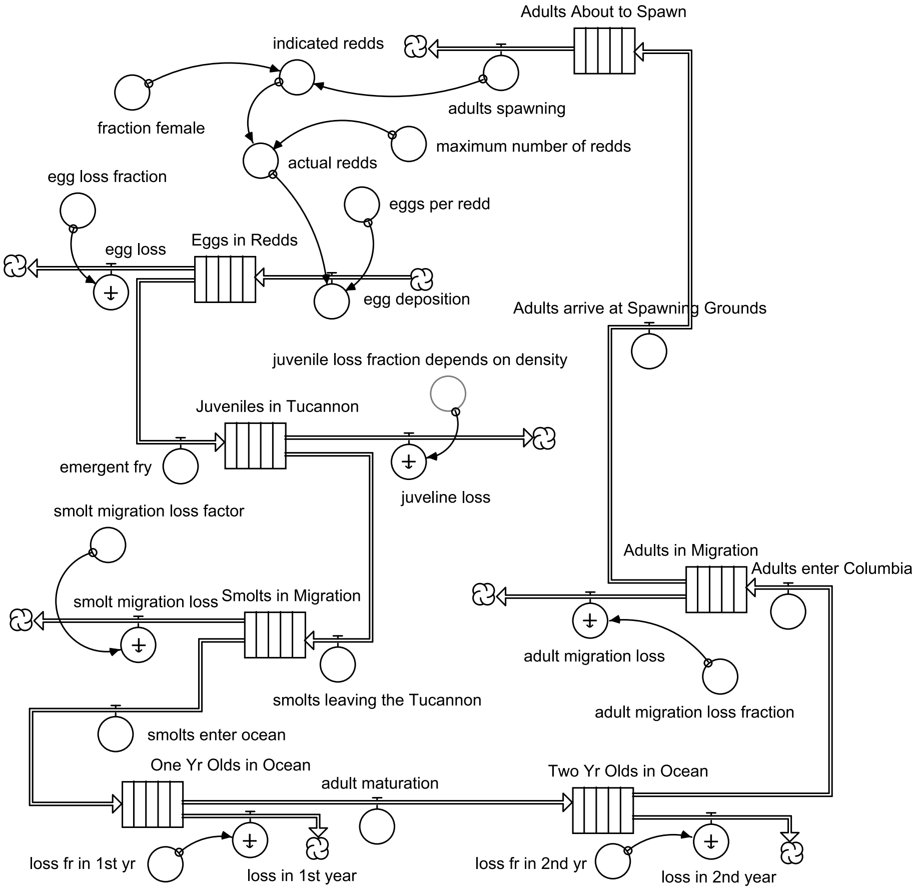
The next example involves the decline in salmon populations in the Snake and Columbia River system of the Pacific Northwest. By the end of the 1990s, the salmon had disappeared from 40% of their historical breeding ranges despite a public and private investment of more than $1 billion. The annual salmon and steelhead runs had dwindled to less than a quarter of the runs from one hundred years ago. Figure 5 shows a system dynamics model one of the salmon runs, the population of Spring Chinook that spawn in the Tucannon River. The river rises in the Blue Mountains of Oregon and flows 50 miles toward the Snake River in Eastern Washington. It is estimated that the river originally supported runs of 20 thousand adults. But the number of returning adults has declined substantially due to many changes in the past sixty years. These changes include agricultural development in the Tucannon watershed, hydro-electric development on the Snake and Columbia, and harvesting in the ocean.

System Dynamics Models of Environment, Energy and Climate Change, Table 1
Inputs to simulate the salmon population under pre-development conditions

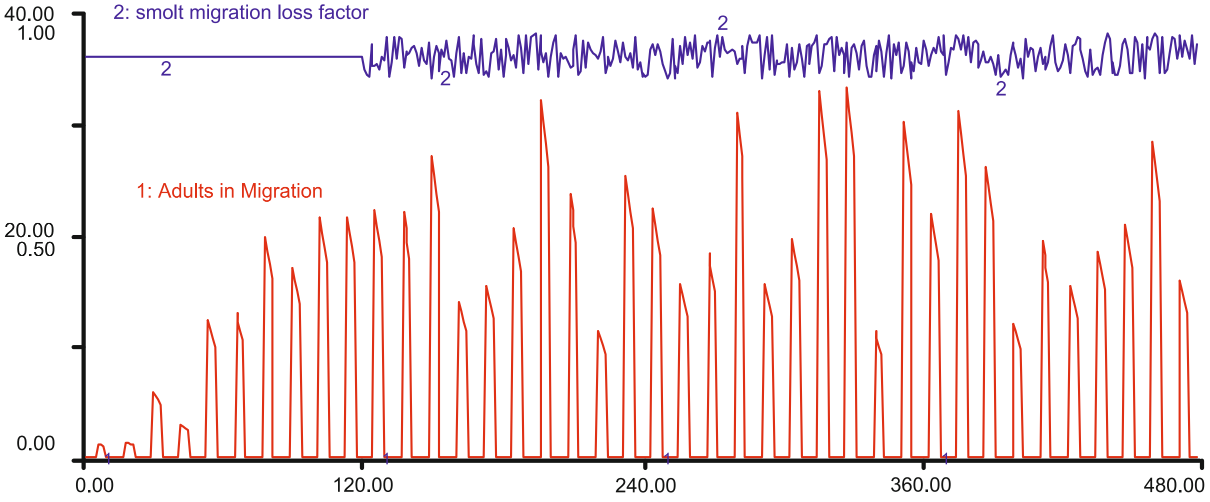
Months in each phase		Population parameters	
Adults ready to spawn	1	fraction female	50%
eggs in redds	6	eggs per redd	3,900
juveniles in Tucannon	12	egg loss fraction	50%
smolts in migration	1	smolt migration loss factor	90%
one yr olds in ocean	12	loss fr for first yr	35%
two yr olds in ocean	12	loss fr for second yr	10%
adults in migration	4	adult migration loss fraction	25%

Each of the stocks in Fig. 5 correspond to a different phase in the salmon life cycle (see Table 1), with a total lifecycle of 48 months. The parameters represent predevelopment conditions, the conditions prior to agricultural development in the Tucannon watershed and hydro-electric development on the Snake and Columbia. Each of these parameters is fixed regardless of the size of the salmon populations. One of the most important variables is the "juvenile loss fraction depends on density." It can be as low as 50% when there are only a few emergent fry each spring. With higher densities, however, juvenile survival becomes more difficult due to crowding in the cool and safe portions of the river.

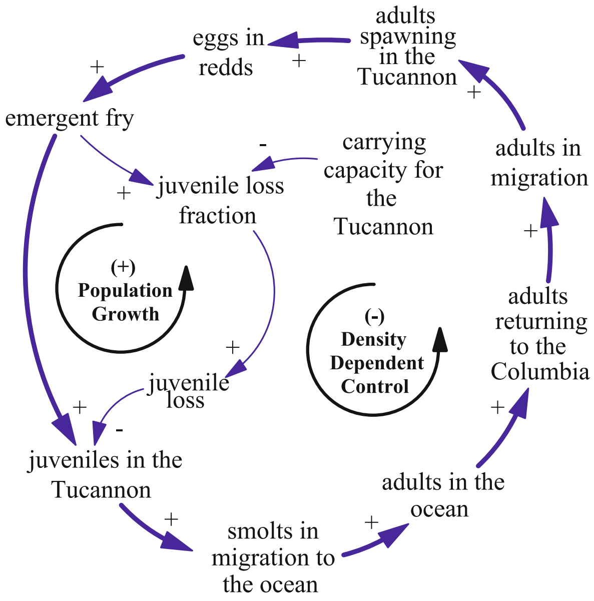
Figure 6 shows the model results over a 480 month period with the population parameters in Table 1. The simulation begins with a small number to see if the population will grow to the 20 thousand adults that were thought to have returned to the river in earlier times. The time graph shows a rapid rise to around 20 thousand adults within the first 120 months of the simulation. The remainder of the simulation tests the population response to variability in environmental conditions, as represented by random variations in the smolt migration loss fraction. (This loss tends to be high in years with low runoff and low in years with high runoff.) Figure 6 confirms that the model simulates the major swings in returning adults due to environmental variability. The runs can vary from a low of ten thousand to a high of thirty thousand.



System Dynamics Models of Environment, Energy and Climate Change, Figure 5
Stella diagram of the model of the salmon life cycle



System Dynamics Models of Environment, Energy and Climate Change, Figure 6
Test of the salmon model with random variations in the smolt migration losses


adults spawning in the Tucannon
eggs in redds
emergent fry
juvenile loss fraction
carrying capacity for the Tucannon
adults in migration
(+) Population Growth
(-) Density Dependent Control
adults returning to the Columbia
juvenile loss
juveniles in the Tucannon
adults in the ocean
smolts in migration to the ocean
+
-

System Dynamics Models of Environment, Energy and Climate Change, Figure 7
Key feedback loops in the salmon model

System dynamics models are especially useful when they help us to understand the key feedbacks in the system. Positive feedback loops are essential to our understanding of rapid, exponential growth; negative feedbacks are essential to our understanding of the controllability of the system. Causal loop diagrams are often used to depict the feedback loops at work in the simulated system. Figure 7 shows an example by emphasizing the most important feedback loops in the salmon model.

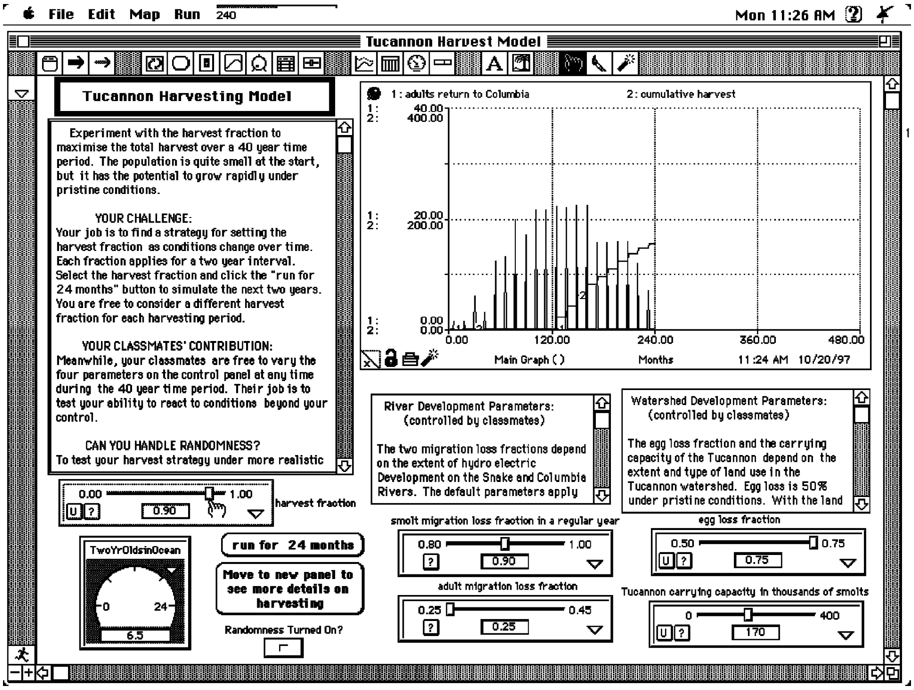
Most readers will immediately recognize the importance of the outer loop which is highlighted by bold arrows in the diagram. Starting near the top, imagine that there are more spawning adults and more eggs in redds. We would then expect to see more emergent fry, more juveniles, more smolts in migration, more salmon in the ocean, more adults entering the Columbia, and a subsequent increase in the number of spawning adults. This is the positive feedback loop that gives the salmon population the opportunity to grow rapidly under favorable conditions.

An equally important feedback works its way around the inner loop in the diagram. If we begin at the top with more spawners, we would expect to see more eggs, more fry and a greater juvenile loss fraction as the fry compete for space in the river. With a higher loss fraction, we expect to see fewer juveniles survive to be smolts, fewer smolts in migration, and fewer adults in the ocean. This means we would see fewer returning adults and less egg deposition. This "density dependent feedback" becomes increasingly strong with larger populations, and it turns out to be crucial to the eventual size of the population. Simulating density dependent feedback is also essential to our understanding of the recovery potential of the salmon population. Suppose, for example, that the salmon experience high losses during the adult migration, This will mean that fewer adults reach the spawning grounds. There will be less egg deposition and fewer emergent fry in the following spring. The new cohort of juveniles will then experience more favorable conditions, and a larger fraction will survive the juvenile stage and migrate to the ocean. The density dependent feedback is crucial to the population's ability to withstand shocks from external conditions.[2]

Figure 8 shows a version of the model to encourage student experimentation with harvesting policies. The information fields instruct the students to work in groups of three with one student playing the role of "the harvest manager". The harvest manager's goal is to achieve a large, sustainable harvest through control of the harvest fraction. The other students are given control of the parameters that describe conditions on the Snake and Columbia and in the Tucannon watershed. These students are encouraged to make major and unpredictable changes to test the instincts of the harvest manager.

Models designed for highly interactive simulations of this kind are sometimes called "management flight simulators" because they serve the same function as actual flight simulators. With a pilot simulator, the trainee takes the controls of an electro-mechanical model and tests his instincts for managing the simulated airplane under difficult conditions. The Tucannon harvesting model provides a similar opportunity for environmental students. They can learn the challenge of managing open access fisheries that are vulnerable to over harvesting and the tragedy of the commons [12]. In this particular exercise, students learn that they can achieve a sustainable harvest under a wide variety of difficult and unpredictable conditions. The key to sustainability is harvest manager's freedom to change the harvest fraction in response to recent trends in number of returning adults. This is an important finding for fishery management because it reveals that the population dynamics are not the main obstacle to sustainability. Rather, unsustainable harvesting is more likely to occur

[2] The shocks could take the form of changes in ocean mortalities, changes in harvesting and changes in the migration mortalities. These shocks are external to the boundary of this model, so one is reminded of Coyle's definition of system dynamics. That is, the model helps us understand how the salmon population could withstand the shocks which fall upon it from the out-side world.



Tucannon Harvesting Model

Experiment with the harvest fraction to maximise the total harvest over a 40 year time period. The population is quite small at the start, but it has the potential to grow rapidly under pristine conditions.

YOUR CHALLENGE:
Your job is to find a strategy for setting the harvest fraction as conditions change over time. Each fraction applies for a two year interval. Select the harvest fraction and click the "run for 24 months" button to simulate the next two years. You are free to consider a different harvest fraction for each harvesting period.

YOUR CLASSMATES' CONTRIBUTION:
Meanwhile, your classmates are free to vary the four parameters on the control panel at any time during the 40 year time period. Their job is to test your ability to react to conditions beyond your control.

CAN YOU HANDLE RANDOMNESS?
To test your harvest strategy under more realistic

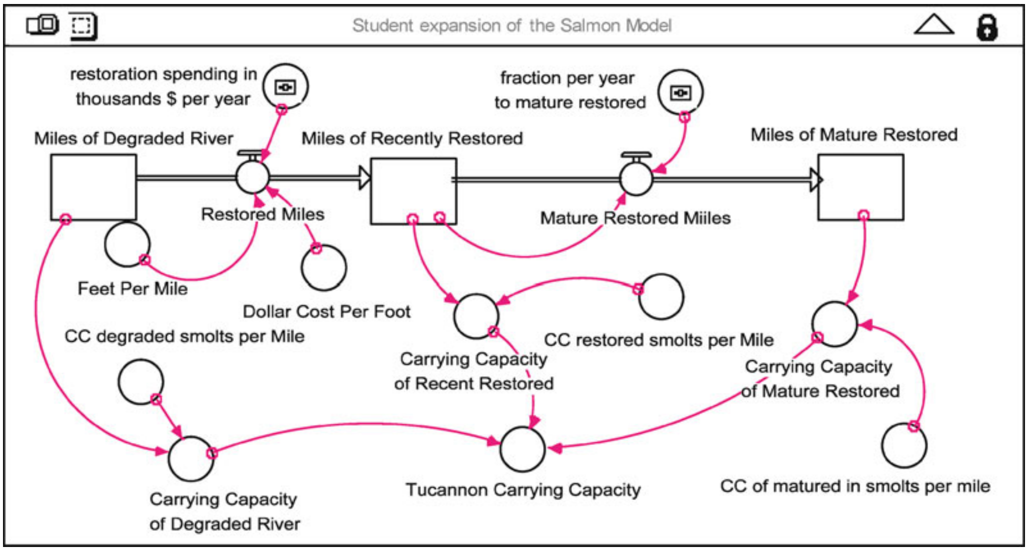
River Development Parameters:
(controlled by classmates)

The two migration loss fractions depend on the extent of hydro electric Development on the Snake and Columbia Rivers. The default parameters apply

Watershed Development Parameters:
(controlled by classmates)

The egg loss fraction and the carrying capacity of the Tucannon depend on the extent and type of land use in the Tucannon watershed. Egg loss is 50% under pristine conditions. With the land

System Dynamics Models of Environment, Energy and Climate Change, Figure 8
Salmon harvesting model to encourage student experimentation


Student expansion of the Salmon Model
restoration spending in thousands $ per year
fraction per year to mature restored
Miles of Degraded River
Miles of Recently Restored
Miles of Mature Restored
Restored Miles
Mature Restored Miiles
Feet Per Mile
Dollar Cost Per Foot
CC degraded smolts per Mile
CC restored smolts per Mile
Carrying Capacity of Recent Restored
Carrying Capacity of Mature Restored
Carrying Capacity of Degraded River
Tucannon Carrying Capacity
CC of matured in smolts per mile

System Dynamics Models of Environment, Energy and Climate Change, Figure 9
Student addition to simulate river restoration

when the managers find it difficult to change the harvest fraction in response to recent trends. This is the fundamental challenge of an open-access fishery.

The salmon model is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation compared to these models. It also provides a launching point for model expansions that can go beyond population biology. Figure 9 shows an example. This is a student expansion to change the carrying capacity from a user input to a variable that responds to the user's river restoration strategy. The student was trained in geomorphology and was an expert on restoring degraded rivers in the west. The Tucannon began the simulation with 25 miles of river in degraded condition and the remaining 25 miles in a mature, fully restored river with a much higher carrying capacity. The new model permits one to experiment with the timing of river restoration spending and to learn the impact on the management of the salmon fishery.

The student's model provides another example of interdisciplinary modeling that aids our understanding of environmental systems. In this particular case, the modeling of river restoration is normally the domain of the geomorphologist. The model of the salmon population is the domain of the population biologist. Their work is often conducted separately, and their models are seldom connected. This is unfortunate as the experts working in their separate domains miss out on the insights that arise when two perspectives are combined within a single model. In the student's case, surprising insights emerged when the combined model was used to study the economic value of the harvesting that could be sustained in the decades following the restoration of the river. To the student's surprise, the new harvesting could "pay back" the entire cost of the river restoration in less than a decade.

Models of Climate Change

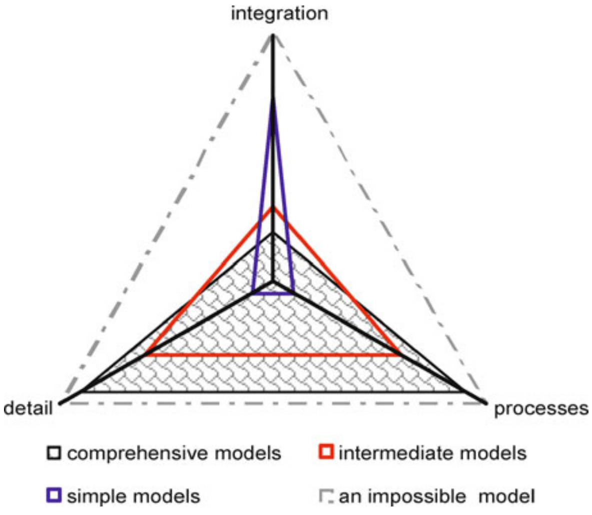
Scientists use a variety of models to keep track of the greenhouse gasses and their impact on the climate. Some of the models combine simulations of the atmosphere, soils, biomass and ocean response to anthropogenic emissions. The more developed models include CO_2, methane, nitrous oxides and other greenhouse gas (GHG) emissions and their changing concentrations in the atmosphere. Claussen [2] classifies climate models as simple, intermediate and comprehensive. The simple models are sometimes called "box models" since they represent the storage in the system by highly aggregated stocks. The parameters are usually selected to match the results from more complicated models. The simple models can be simulated faster on the computer, and the results are easier to interpret. This makes them valuable for sensitivity studies and in scenario analysis [13].

The comprehensive models are maintained by large research centers, such as the Hadley Center in the UK. The term "comprehensive" refers to the goal of capturing all the important processes and simulating them in a highly detailed manner. The models are sometimes called GCMs (general circulation models). They can be used to describe circulation in the atmosphere or the ocean. Some simulate both the ocean and atmospheric circulation in a simultaneous, interacting fashion. They are said to be coupled general circulation models (CGCMs) and are considered to be the "most comprehensive" of the models available [2]. They are particularly useful when a high spatial resolution is required. However, a disadvantage of the CGCMs is that only a limited number of multi-decadal experiments can be performed even when using the most powerful computers.

Intermediate models help scientists bridge the gap between the simple and the comprehensive models. Claussen [2] describes eleven models of intermediate complexity. These models aim to "preserve the geographic integrity of the Earth system" while still providing the opportunity for multiple simulations to "explore the parameter space with some completeness. Thus, they are more suitable for assessing uncertainty". Figure 10 characterizes the different categories of models based on their relative emphasis on:

- number of processes (right axis)
- detailed treatment of the each process (left axis), and the
- extent of integration among the different processes (top axis).

Regardless of the methodology, climate modeling teams must make some judgments on where to concentrate their attention. No model can achieve maximum performance along all three dimensions. (Figure 10 uses the dashed lines



System Dynamics Models of Environment, Energy and Climate Change, Figure 10
Classification of climate models

to draw our attention to the impossible task of doing every thing within a single model.)

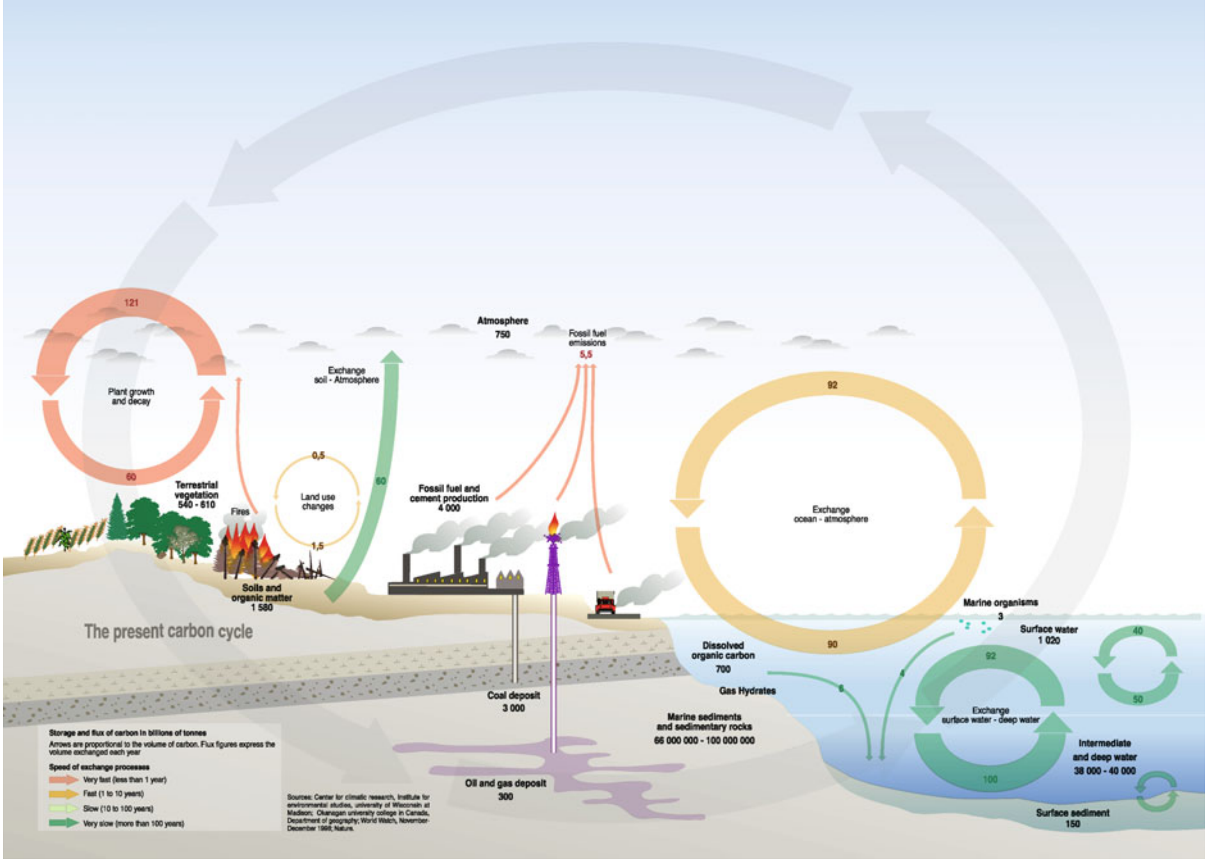
The comprehensive models strive to simulate as many processes as possible with a high degree of detail. This approach provides greater realism, but the models often fail to simulate the key feedback loops the link that atmospheric system with the terrestrial and oceanic systems. (An example is the feedback between CO_2 emissions, temperatures and the decomposition of soil carbon. If higher temperatures lead to accelerated decomposition, the soils could change from a net sink to a net source of carbon [15].) The simple models sacrifice detail and the number of processes in order to focus on the feedback effects between the processes. Using Claussen's terminology, one would say that such models aim for a high degree of "integration". However, the increased integration is achieved by limiting the number of processes and the degree of detail in representing each of the processes.

System dynamics has been used in a few applications to climate change. These applications fit in the category of simple models whose goal is to provide a highly integrated representation of the system. Two examples are described here; both deal with the complexities of the global carbon cycle.

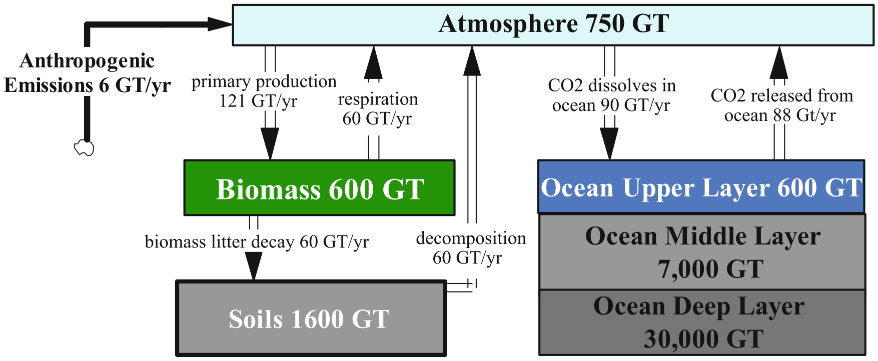
System Dynamics Models of the Carbon Cycle

Figures 11 and 12 depict the global carbon cycle. Figure 11 shows the carbon flows in a visual manner. Figure 12 uses the Vensim stock and flow icons to summarize carbon storage and flux in the current system. The storage is measured in GT, gigatons of carbon, (where carbon is the C in CO_2). The flows are in GT/year of carbon with values rounded off for clarity.

The left side of Fig. 12 shows the flows to the terrestrial system. The primary production removes 121 GT/yr from the atmosphere. This outflow exceeds the return flows by 1 GT/year. This imbalance suggests that around 1 GT of carbon is added to the stocks of biomass and soil each year. So the carbon stored in the terrestrial system would


121
Plant growth and decay
60
Terrestrial vegetation 540 - 610
Fires
Exchange soil - Atmosphere
60
Land use changes
0,5
1,5
Atmosphere 750
Fossil fuel emissions 5,5
Fossil fuel and cement production 4 000
Soils and organic matter 1 580
The present carbon cycle
Coal deposit 3 000
Oil and gas deposit 300
Storage and flux of carbon in billions of tonnes
Arrows are proportional to the volume of carbon. Flux figures express the volume exchanged each year
Speed of exchange processes
Very fast (less than 1 year)
Fast (1 to 10 years)
Slow (10 to 100 years)
Very slow (more than 100 years)
Sources: Center for climatic research, Institute for environmental studies, university of Wisconsin at Madison; Okanagan university college in Canada, Department of geography; World Watch, November-December 1998; Nature.
92
Exchange ocean - atmosphere
90
Marine organisms 3
Surface water 1 020
40
50
92
Dissolved organic carbon 700
Gas Hydrates
6
4
Exchange surface water - deep water
100
Marine sediments and sedimentary rocks 66 000 000 - 100 000 000
Intermediate and deep water 38 000 - 40 000
Surface sediment 150

System Dynamics Models of Environment, Energy and Climate Change, Figure 11
The global carbon cycle. (Source: United Nations Environmental Program (UNEP) http://www.unep.org/)

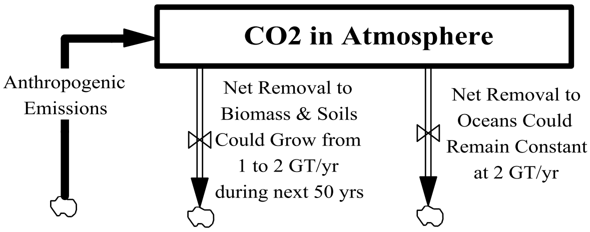

Atmosphere 750 GT
Anthropogenic Emissions 6 GT/yr
primary production 121 GT/yr
respiration 60 GT/yr
CO2 dissolves in ocean 90 GT/yr
CO2 released from ocean 88 Gt/yr
Biomass 600 GT
Ocean Upper Layer 600 GT
biomass litter decay 60 GT/yr
decomposition 60 GT/yr
Ocean Middle Layer 7,000 GT
Soils 1600 GT
Ocean Deep Layer 30,000 GT

System Dynamics Models of Environment, Energy and Climate Change, Figure 12
Diagram of the stocks and flows in the carbon cycle

grow over time (perhaps due to extensive reforestation of previously cleared land.) The right side of Fig. 12 shows the flows from the atmosphere to the ocean. The CO_2 dissolved in the ocean each year exceeds the annual release back to the atmosphere by 2 GT. The total, net-flow out of the atmosphere is 3 GT/year which means that natural processes are acting to negate approximately half of the current anthropogenic load.

As the use of fossil fuels grows over time, the anthropogenic load will increase. But scientists do not think that natural processes can continue to negate 50% of an ever increasing anthropogenic load. On the terrestrial side of the system, there are limits on the net flow associated with reforestation of previously cleared land. And there are limits to the carbon sequestration in plants and soils due to nitrogen constraints. On the ocean side of the system, the current absorption of 2 GT/year is already sufficiently high to disrupt the chemistry of the ocean's upper layer. Higher CO_2 can reduce the concentration of carbonate, the ocean's main buffering agent, thus affecting the ocean's ability to absorb CO_2 over long time periods.

Almost of the intermediate and comprehensive climate models may be used to estimate CO_2 accumulation in the atmosphere in the future. For this article, it is useful to draw on the mean estimate published in *Climatic Change* by Webster [23]. He used the climate model developed at the Massachusetts Institute of Technology, one of the eleven models of "intermediate complexity" in the review by Claussen [2]. The model began the simulation in the year 2000 with an atmospheric CO_2 concentration of 350 parts per million (ppm). (This concentration corresponds to around 750 GT of carbon in the atmosphere.) The mean projection assumed that anthropogenic emissions would grow to around 19 GT/year by 2100. The mean projection of atmospheric CO_2 was around 700 ppm by 2100. The amount of CO_2 in the atmosphere would be twice as high at the end of the century.


CO2 in Atmosphere
Anthropogenic Emissions
Net Removal to Biomass & Soils Could Grow from 1 to 2 GT/yr during next 50 yrs
Net Removal to Oceans Could Remain Constant at 2 GT/yr

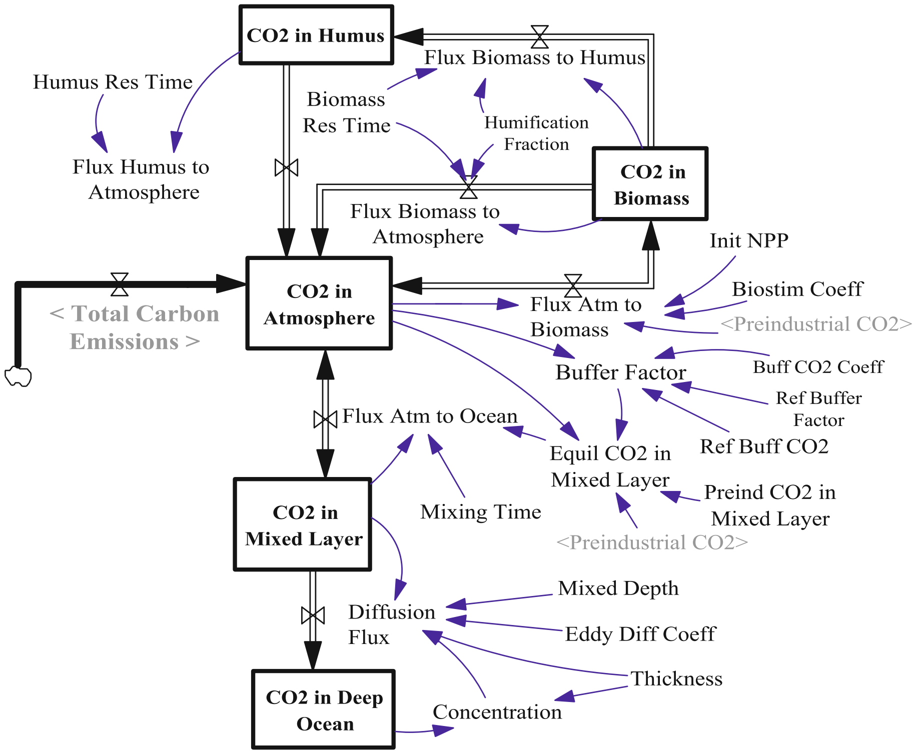
System Dynamics Models of Environment, Energy and Climate Change, Figure 13
Simple model to understand accumulation of CO_2 in the atmosphere

Figure 13 shows the simplest possible model to explain the doubling of atmospheric CO_2. The stock accumulates the effect of three flows, each of which is specified by the user. Anthropogenic emissions are set to match Webster's assumption. They grow to 19 GT/year by the end of the century. Net removal to oceans is assumed to remain constant at 2 GT/year for the reasons given previously. Net removal to biomass and soils is then subject to experimentation to allow this simple model to match Webster's results. A close match is provided if the net removal increases from 1 to 2 GT/year during the first half of the century and then remains at 2 GT/year for the next fifty years. With these assumptions, the CO_2 in the atmosphere would double from 750 to 1500 GT during the century. This means that the atmospheric concentration would double from 350 to 700 ppm, the same result published by Webster [23].

The model in Fig. 13 is no more than an accumulator. This is the simplest of possible models to add insight on the dynamics of CO_2 accumulation in the atmosphere. It includes a single stock and only three flows, with all of the flows specified by the user. There are no feedback relationships which are normally at the core of system dynamics models. This extreme simplification is intended to make the point that simple models may provide perspective on the dynamics of a system. In this case, a simple accumulator can teach one about the sluggish response of atmospheric CO_2 in the wake of reductions in the anthropogenic emissions. As an example, suppose carbon policies were to succeed in cutting global emissions dramatically in the year 2050. By this year, emissions would have reached 10 GT/yr, so the supposed policy would reduce emissions to 5 GT/yr. What might then happen to CO_2 concentrations in the atmosphere for the remainder of the century? Experiments with highly educated adults [21] suggest that some subjects would answer this question with "pattern matching" reasoning. For example, if emissions are cut in half, it might make sense that CO_2 concentrations would be cut in half as well. But pattern matching leads one astray since the accumulation of CO_2 in the atmosphere responds to the total effect of the flows in Fig. 13. Were anthropogenic emissions to be reduced to 5 GT/year and net removals were to remain at 4 GT/year, the CO_2 concentration would continue to grow, and atmospheric CO_2 would reach 470 ppm by the end of the century.

The model in Fig. 13 is an extreme example to make a point about the usefulness of simple models. The next example is by Fiddaman [6]. It was selected as illustrative of the type of model that would emerge after a system dynamics study. Figure 14 shows the view of the carbon cycle, one of 30 views in the model. The model simulates the climate system within a larger system that includes growth in human population, growth in the economy, and changes in the production of energy. The model was organized conceptually as nine interacting sectors with a high degree of coupling between the energy, economic and the climate sectors.

Fiddaman focused on policy making, particularly the best way to put a price on carbon. In the current debate, this question comes down to a choice between a carbon tax and a carbon market. His simulations add support to those who argue that the carbon tax is the preferred method of putting a price on carbon. The simulations also provide another example of the usefulness of system dynamics models that cross disciplinary boundaries. By representing the economy, the energy system and the climate system



System Dynamics Models of Environment, Energy and Climate Change, Figure 14
Representation of the carbon cycle in the model by Fiddaman [6]

within a single, tightly coupled model, he provides another example of the power of system dynamics to promote interdisciplinary exploration of complex problems.

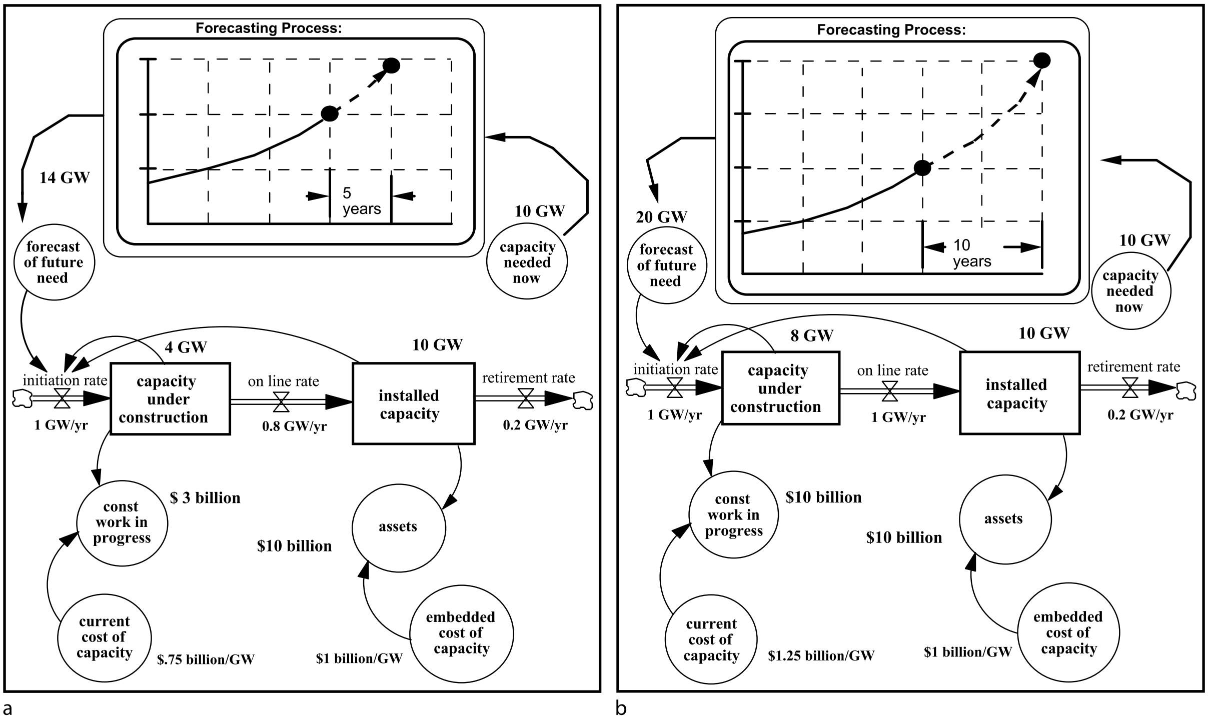
System dynamics has also been applied to a wide variety of energy problems [1,7]. Indeed, a key word frequency count in 2004 revealed nearly 400 energy entries in the System dynamics bibliography [11]. Many of these applications deal with the electric power industry, and I have selected two electric studies to illustrate the usefulness of the approach. The first involves the regulatory and financial challenges of the investor owned electric utilities in the United States.

Lessons from the Regulated Power Industry in the 1970s

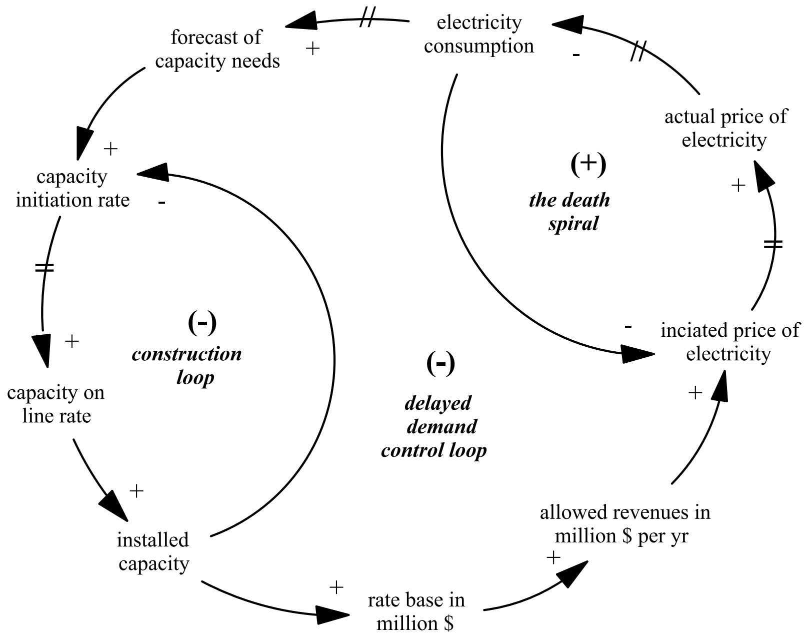
The 1970s was a difficult decade for the regulated power companies in the United States. The price of oil and gas was increasing rapidly, and the power companies were frequently calling on their regulators to increase retail rates to cover the growing cost of fuel. The demand for electricity had been growing rapidly during previous decades, often at 7 %/year. At this rate, the demand doubled every decade, and the power companies faced the challenge of doubling the amount of generating capacity to ensure that demand would be satisfied. The power companies dealt with this challenge in previous decades by building ever larger power plants (whose unit construction costs declined due to economies of scale). But the economies of scale were exhausted by the 1970s, and the power companies found themselves with less internal funds and poor financial indicators. Utilities worried that the construction of new power plants would not keep pace with demand, and the newspapers warned of curtailments and blackouts.

Figure 15 puts the financial problems in perspective by showing the forecasting, planning and construction processes. The side by side charts allows one to compare the difficult conditions of the 1970s with conditions in previous decades. Figure 15a shows the situation in the 1950s and 1960s. Construction lead times were around 5 years, so forecasts would extend 5 years into the future. Given the costs at the time, the power company would need to finance $3 billion in construction. This was a substantial, but manageable task for a company with $10 billion in assets.

Figure 15b shows the dramatic change in the 1970s. Construction lead times had grown to around 10 years, and construction costs had increased as well. The power



System Dynamics Models of Environment, Energy and Climate Change, Figure 15
a The electric utility's financial challenge during the 1950s and 1960s. b The electric utility's financial challenge during the 1970s



System Dynamics Models of Environment, Energy and Climate Change, Figure 16
Key feedbacks and delays faced by power companies in the 1970s

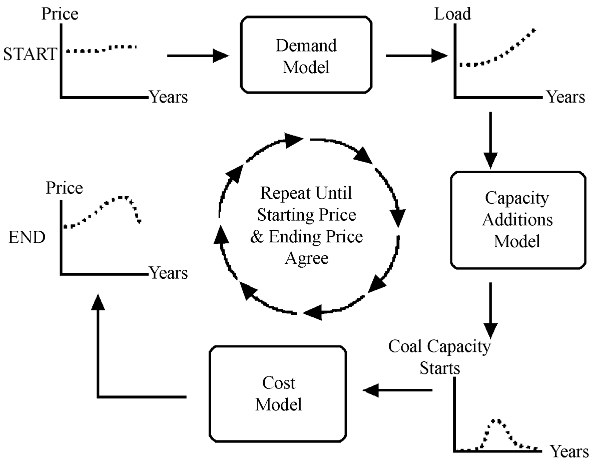
company faced the challenge of financing $10 billion in construction with an asset base of $10 billion. The utility executives turned to the regulators for help. They asked for higher electricity rates in order to increase annual revenues and improve their ability to attract external financing. The regulators responded with substantial rate increases, but they began to wonder whether further rate increases would pose a problem with consumer demand. If consumers were to lower electricity consumption, the utility would have less sales and less revenues. The executives might then be forced to request another round of rate increases. Regulators wondered if they were setting loose a "death spiral" of ever increasing rates, declining sales and inadequate financing.

Figure 16 puts the problem in perspective by showing the consumer response to higher electricity rates along side of the other key feedback loops in the system. Higher electricity rates do pose the problem which came to be called "the death spiral". But the death spiral does not act in isolation. Figure 16 reminds us that higher rates lead to lower consumption and to a subsequent reduction in the demand forecast and in construction. After delays for the new power plants to come on line, the power companies experiences a reduction in its "rate base" and the "allowed revenues". When the causal relationships are traced around the outer loop, one sees a negative feedback loop that could act to stabilize the situation. The problem, however, is that the delays around the outer loop are substantially longer than the delay for the death spiral.

The utility companies financial challenge was the subject of several system dynamics studies in the 1970s and 1980s [7]. The studies revealed that the downward spiral could pose difficult problems, especially if consumers reacted quickly while utilities were stuck with long-lead time, capital intensive power plants under construction. The studies showed that utility executives needed to do more than rely on regulators to grant rate increases; they needed to take steps on their own to soften the impact of the death spiral. The best strategy was to shift the investments to technologies with shorter lead times. (As an example, a power company in coal region would do better to switch from large to smaller coal plants because of the small plants' shorter lead time.) The studies also revealed that the company's financial situation would improve markedly with slower growth in demand. By the late 1970s and early 1980s, many power companies began to provide direct financial incentives to their customers to slow the growth in demand. System dynamics studies showed that the company-sponsored efficiency programs would be beneficial to the both the customers (lower electric bills) and to the power companies (improved financial performance).

An essential feature of the utility modeling was the inclusion of power operations along side of consumer behavior, company forecasting, power plant construction, regulatory decision making and company financing. This interdisciplinary approach is common within the system dynamics community because practitioners believe that insights will emerge from simulating the key feedback loops. (This belief leads one to follow the cause and effect connections around the key loops regardless of the disciplinary boundaries that are crossed along the way.) This approach contrasts strongly with the customary modeling framework of large power companies who were not familiar with system dynamics. Their approach was to assign models to different departments (i. e., operations, accounting and forecasting) and string the models together to provide a view of the entire corporation over the long-term planning interval.

Figure 17 shows what can happen when models within separate departments are strung together. A large corporation might use 30 models, but this diagram makes the point by describing three models. The analysis would begin with an assumption on future electricity prices over the 20-year interval. These are needed to prepare a forecast of the growth in electricity load. The forecast is then given to the planning department which may run a variety of models to select the number power plants to construct in the future. The construction results are then handed to the accounting and rate making departments to prepare a forecast of electricity prices. When the company finally completes the many calculations, the prices that emerge may not agree with the prices that were assumed at the start. The company must then choose whether to ignore the contradiction or to repeat the entire process with a new estimate of the prices at the top of the diagram. This was not an easy choice. Ignoring the price discrepancy was problematic because it was equivalent to ignoring the "death spiral," one of the foremost problems of the 1970s. Repeating the analysis was also problematic. The new round of calculations would be time consuming, and there was no guarantee that consistent results would be obtained at the end of the next iteration.



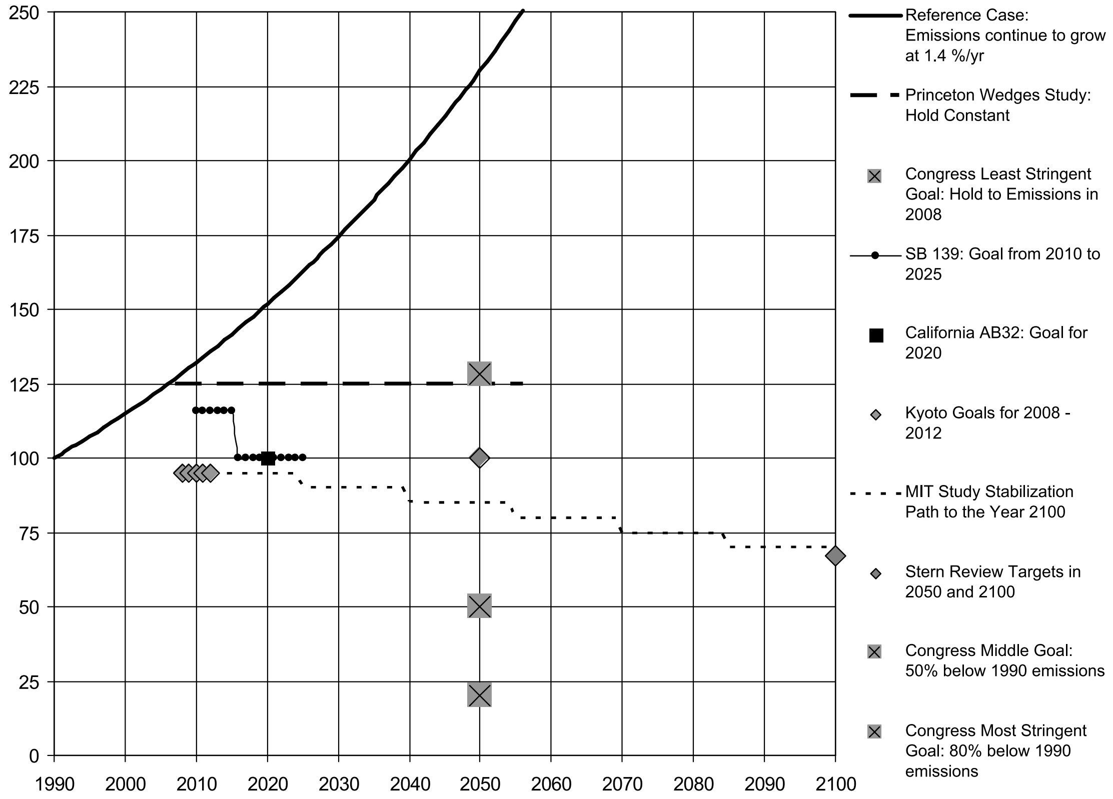
System Dynamics Models of Environment, Energy and Climate Change, Figure 17
The iterative approach often used by large power companies in the 1970s

The power companies' dilemma from the 1970s is described here to make an important point about the usefulness of system dynamics. System dynamics modeling is ideally suited for the analysis of dynamic problems that require a feedback perspective. The method allows one to "close the loop", as long as one is willing to cross the necessarily disciplinary boundaries. In contrast, other modeling methods are likely to be extremely time consuming or fall short in simulating the key feedbacks that tie the system together.

Simulating the Power Industry Response to a Carbon Market

The world is getting warmer, both in the atmosphere and in the oceans. The clearest and most emphatic description of global warming was issued by the intergovernmental panel on climate change (IPCC) in February of 2007. Their summary for policymakers (p. 4 in [14]) reported that the "Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice and rising global mean sea level". The IPCC concluded that "most of the observed increase is very likely due to the observed increase in anthropogenic greenhouse gas concentrations". As a consequence of the IPCC and other warnings, policymakers around the world are calling for massive reductions in CO_2 and other greenhouse gas (GHG) emissions to reduce the risks of global warming.

Figure 18 summarizes some of the targets for emission reductions that have been adopted or proposed around the world. In many cases, the targets are specified relative to a country's emissions in the year 1990. So, for ease of comparison, the chart uses 100 to denote emissions in the year 1990. Emissions have been growing at around 1.4%/year. The upward curve shows the future emissions if this trend continues: emissions would reach 200 by 2040 and 400 by 2090. The chart shows the great differences in the stringency of the targets. Some call for holding emissions constant; others call for dramatic reductions over time. Some targets apply to the next two decades; many extend to the year 2050; and some extend to the year 2100. However,

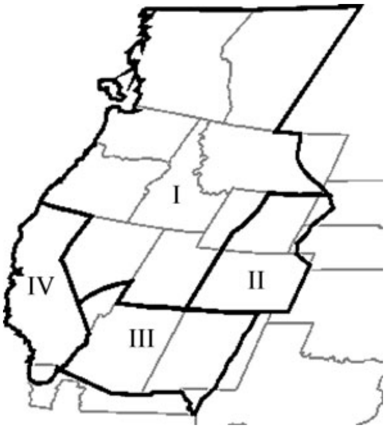


System Dynamics Models of Environment, Energy and Climate Change, Figure 18
Comparison of goals for emissions (100 on the vertical axis represent emissions in the year 1990)

when compared to the upward trend, all targets require major reductions relative to business as usual.

The targets from the Kyoto treaty are probably the best known of the goals in Fig. 18. The treaty became effective in February of 2005 and called for the Annex I countries to reduce emissions, on average, by 5% below 1990 emissions by the year 2008 and to maintain this limit through 2012. The extension of the Kyoto protocol beyond 2012 is the subject of ongoing discussions. The solid line from 2010 to 2050 represents the "stabilization path" used in the climate modeling by Webster [23]. The limit on emissions was imposed in modeling calculations designed to stabilize atmospheric CO_2 at 550 ppmv or lower. The scenario assumed that the Kyoto emissions caps are adopted by all countries by 2010. The policy assumed that the caps would be extended and then further lowered by 5% every 15 years. By the end of the century, the emissions would be 35% below the value in 1990.

This article concentrates on Senate Bill 139, The Climate Stewardship Act of 2003. Figure 19 shows the S139 targets over the interval from 2010 to 2025. The bill called for an initial cap on emissions from 2010 to 2016. The cap would be reduced to a more challenging level in 2016, when the goal was to limit emissions to no more than the emissions from 1990. S139 was introduced by Senators McCain and Lieberman in January of 2003. It did



System Dynamics Models of Environment, Energy and Climate Change, Figure 19
Map of the western electricity system

not pass, but it was the subject of several studies including a highly detailed study by the Energy Information Administration [5]. The EIA used a wide variety of models to search for the carbon market prices that would induce industries to lower emissions to come into compliance with the cap. The carbon prices were estimated at $22 per metric ton of CO_2 when the market was to open in 2010. They were projected to grow to $60 by the year 2025.

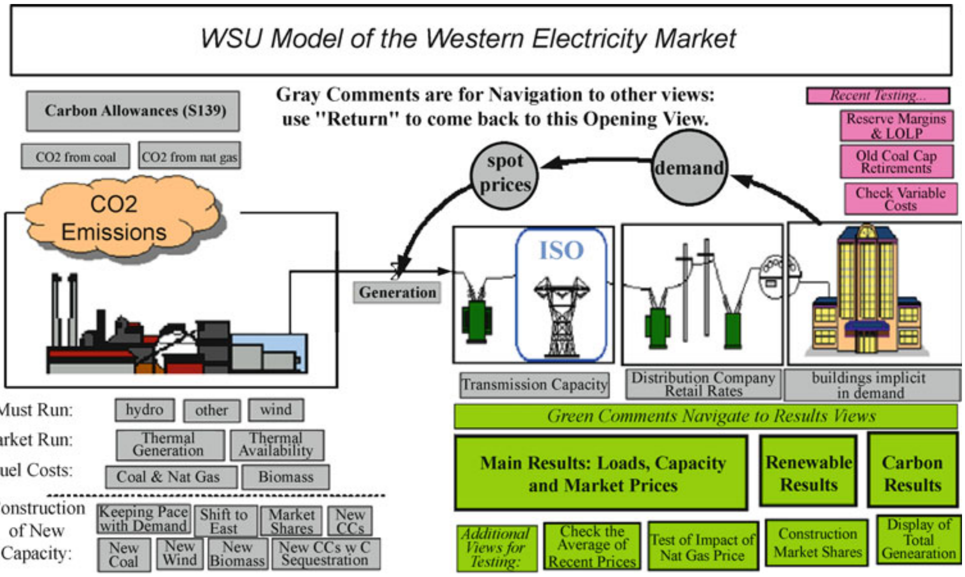
The EIA study showed that the electric power sector would lead the way in reducing emissions. By the year 2025, power sector emissions would be reduced 75% below the reference case. This reduction was far beyond the reductions to be achieved by other sectors of the economy. This dramatic response was possible given the large use of coal in power generation and the power industry's wide range of choices for cleaner generation.

A system dynamics study of S139 was conducted at Washington State University (WSU) to learn if S139 could lead to similar reductions in the west. Electricity generation in the western system is provided in a large, interconnected power system shown in Fig. 19. This region has considerably more hydro resources, and it makes less use of coal-fired generation than the nation as a whole. The goal was to learn if dramatic reductions in CO_2 emissions could be possible in the west and to learn if they could be achieved with generating technologies that are commercially available today.

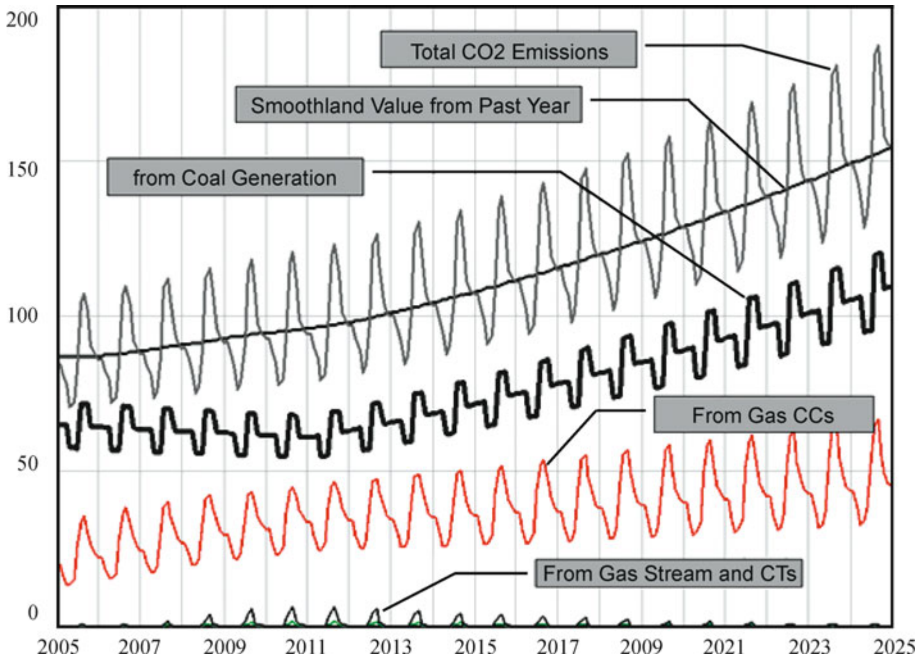
The opening view of the WSU model is shown in Fig. 20. The model deals with generation, transmission and distribution to end use customers, with price feedback on the demand for electricity. The model is much larger than the textbook models described earlier in this article. Fifty views are required to show the all the diagrams and the simulation results. The opening view serves as a central hub to connect with all the other views.

The opening view uses Vensim's comment icons to draw attention to the CO_2 emissions in the model. The emissions arise mainly from coal-fired power plants, as shown in Fig. 21. A smaller, but still significant fraction of the emissions is caused by burning natural gas in combined cycle power plants. Total emissions vary with the seasons of the year, with the peak normally appearing in the summer when almost all of the fossil-fueled plants are needed to satisfy peak demand. The base case shows annual emissions growing by over 75% by the year 2025.

A major challenge for the system dynamics model is representing power flows across a transmission grid. Finding the flows on each transmission line and the prices in each area is difficult with the standard tools of system dynamics. It simply doesn't make sense to represent the power flows with a combination of stocks, flows and feedback processes to explain the flows. It makes more sense to calculate the flows and prices using traditional power systems methods, as explained by Dimitrovski [4]. The power flows were estimated using an algebraic approach which power engineers label as a reduced version of a direct-current optimal power flow calculation. The solution to the algebraic constraints were developed with the Matlab soft-


WSU Model of the Western Electricity Market
Carbon Allowances (S139)
Gray Comments are for Navigation to other views:
use "Return" to come back to this Opening View.
Recent Testing...
Reserve Margins & LOLP
Old Coal Cap Retirements
Check Variable Costs
CO2 from coal
CO2 from nat gas
CO2 Emissions
spot prices
demand
Generation
ISO
Transmission Capacity
Distribution Company Retail Rates
buildings implicit in demand
Green Comments Navigate to Results Views
Main Results: Loads, Capacity and Market Prices
Renewable Results
Carbon Results
Must Run:
hydro
other
wind
Market Run:
Thermal Generation
Thermal Availability
Fuel Costs:
Coal & Nat Gas
Biomass
Construction of New Capacity:
Keeping Pace with Demand
Shift to East
Market Shares
New CCs
New Coal
New Wind
New Biomass
New CCs w C Sequestration
Additional Views for Testing:
Check the Average of Recent Prices
Test of Impact of Nat Gas Price
Construction Market Shares
Display of Total Genearation

System Dynamics Models of Environment, Energy and Climate Change, Figure 20
Opening view of the model of the western electricity system



System Dynamics Models of Environment, Energy and Climate Change, Figure 21
Annual emissions in a base case simulation (annual emissions are in million metric tons of carbon)

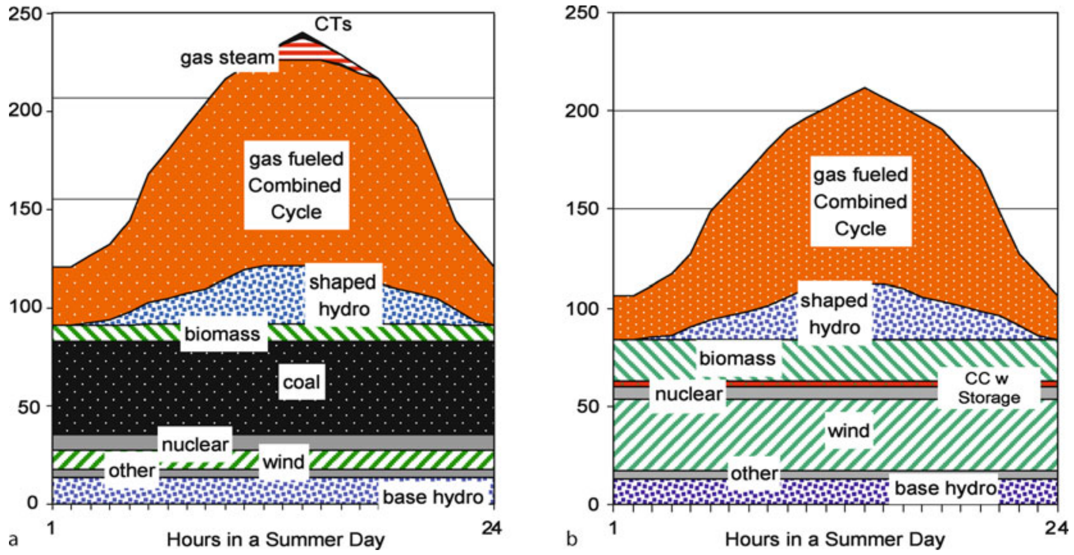
ware and then transferred to user-defined functions to operate within the Vensim software. The Vensim simulations were set to run over twenty years with time in months. (A typical simulation required 240 months with changes during a typical day handled by carrying along separate calculations for each of 24 h in a typical day.) These are extensive calculations compared to many system dynamics models, so there was concern that we would lose the rapid simulation speed that helps to promote interactive exploration and model testing. The important methodological accomplishment of this project was the inclusion of network and hourly results within a long-term model without losing the rapid simulation response that encourages users to experiment with the model.

One of the model experiments called for a new simulation with carbon prices set to follow the \$20 to \$60 trajectory projected by the EIA for S139. These prices were specified as a user input, and the model responded with a change in both short-term operations and long-term investments. The important result was a 75% reduction in CO_2 emissions by the end of the simulation. This dramatic reduction corresponds almost exactly to the EIA estimate of CO_2 reduction for the power industry in the entire US.

Figure 22 helps one understand how CO_2 emissions could be reduced by such a large amount. These diagrams show the operation of generating units across the Western US and Canada for a typical day in the summer of the final year of the simulation. Figure 22a shows the reference case; Figure 22b shows the case with S139. The side by side comparison helps one visualize the change in system operation. A comparison of the peak loads shows that the demand for electricity would be reduced. The reduction is 9%, which is due entirely to the consumers' reaction to higher retail electric prices over time.

Figure 22b shows large contributions from wind and biomass generation. Wind generation is carbon free, and biomass generation is judged to be carbon neutral, so these generating units make an important contribution by the end of the simulation. Both of these generating technologies are competitive with today's fuel prices and tax credits. The model includes combined cycle gas generation equipped with carbon capture and storage, a technology that is not commercially available today. The model assumes that advances in carbon sequestration over the next two decades would allow this technology to capture a small share of investment near the end of the simulation. By the year 2025, the combined cycle plants with sequestration equipment would provide 2% of the generation.

The most important observation from Fig. 22 is the complete elimination of coal-fired generation in the S139 case. Coal-fired units are shown to operate in a base load



System Dynamics Models of Environment, Energy and Climate Change, Figure 22
a Projected generation for a peak summer day in 2024 in the reference case. b Projected generation for a peak summer day in 2024 in the S139 case

mode in the reference simulation. They provide around 28% of the annual generation, but they account for around two/thirds of the CO_2 emissions in the western system. The carbon prices from S139 make investment in new coal-fired capacity unprofitable at the very start of the simulated market in 2010. As the carbon prices increase, utilities to cut back on coal-fired generation and compensate with increased generation from gas-fired CC capacity. In the simulations reported here, this fuel switching would push the coal units into the difficult position of operating fewer and fewer hours in a day. Eventually this short duration operation is no longer feasible, and coal generation is eliminated completely by the end of the simulation.

The WSU study of the western electric system was selected as the concluding example because of its novel treatment of network flows inside a system dynamics model [4]. The model is also interesting for its treatment of daily price changes within a long-term model. (Such changes are important to the simulation of revenues in the wholesale market.) From a policy perspective, the study confirms previous modeling of the pivotal role of the electric power industry in responding to carbon markets. The study indicated that the western electricity system could achieve dramatic reductions in CO_2 emissions within 15 years after the opening of a carbon market, and it could do so with technologies that are commercially available today [8].

Conditions for Effective Interdisciplinary Modeling

All of the applications demonstrate the usefulness of system dynamics in promoting interdisciplinary modeling. The article concludes with comments on the level of effort and the conditions needed for effective, interdisciplinary modeling.

The examples in this article differ substantially in the level of effort required, from several weeks for the classroom examples to several years for the energy studies. The textbook examples involved student expansions of models of Mono Lake and the Tucannon salmon. The expansions were completed by undergraduate students in projects lasting two or three weeks. The key was the students' previous education (classes from many different departments) and their receptiveness to an interdisciplinary approach.

Fiddaman's model of the climate and energy system [6] was a more ambitious exercise, requiring several years of effort as part of his doctoral research. Bringing multi-year interdisciplinary modeling projects to a successful conclusion requires one to invest the time to master several disciplines and to maintain a belief that there are potential insights at the end of the effort.

The electric power industry examples were also ambitious projects that required several years of effort. The modeling of the western electricity system was a four-year project with support from the National Science Foundation. The long research period was crucial for it allowed the researchers from power systems engineering, system dynamics and environmental science to take the time to learn from one another. The modeling of the electric company problems in the 1970s was also spread over several years of effort. The success of this modeling was aided by utility planners, managers and modelers who were looking for a systems view of their agency and its problems. They saw system dynamics as a way to tie existing ideas together within an integrated portrayal of their system. Their existing ideas were implemented in models maintained by separate functional areas (i. e., forecasting, accounting, operations). The existing models often provided a foundation for the system dynamics models (i. e., in the same way that the comprehensive climate models in Fig. 11 provide support for the development of the more integrated models). The key to effective, interdisciplinary modeling within such large organizations is support from a client with a strong interest in learning and with managerial responsibility for the larger system.

Future Directions

This article concludes with future directions for system dynamics applications to climate change. People often talk of mitigation and adaptation. Mitigation refers to the challenge of lowering greenhouse gas emissions to avoid dangerous anthropogenic interference with the climate system. Adaptation refers to the challenge of living in a changing world.

Mitigation: The challenge of lowering CO_2 and other GHG emissions is the fundamental challenge of the coming century. The next two decades will probably see various forms of carbon markets, and system dynamics can aid in learning about market design. It is important that we learn how to make these markets work well. And if they don't work well, it's important to speed the transition to a carbon tax policy with better prospects for success. System dynamics can aid in learning about markets, especially if it is coupled with simulating gaming to allow market participants and regulators to "experience" and better understand market dynamics.

Adaptation: The world will continue to warm, and sea levels will continue to rise. These trends will dominate the first half of this century even with major reductions in CO_2 emissions. These and other climate changes will bring a wide variety of problems for management of water resource, public health planning, control of invasive species, preservation of endangered species, control of wildfire, and coastal zone management, just to name a few. Our understanding of the adaptation challenges can be improved through system dynamics modeling. The prospects for insight are best if the models provide an interdisciplinary perspective on adapting to a changing world.

Bibliography

Primary Literature

1. Bunn D, Larsen E (1997) Systems modelling for energy policy. Wiley, Chichester
2. Claussen M et al (2002) Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. Climate Dyn 18:579–586
3. Coyle G (1977) Management system dynamics. Wiley, Chichester
4. Dimitrovski A, Ford A, Tomsovic K (2007) An interdisciplinary approach to long term modeling for power system expansion. Int J Crit Infrastruct 3(1–2):235–264
5. EIA (2003) United States Department of Energy, Energy Information Administration, Analysis of S139, the Climate Stewardship Act of 2003
6. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–264
7. Ford A (1999) Modeling the environment. Island Press, Washington
8. Ford A (2008) Simulation scenarios for rapid reduction in carbon dioxide emissions in the western electricity system. Energy Policy 36:443–455
9. Forrester J (1961) Industrial dynamics. Pegasus Communications
10. Forrester J (2000) From the ranch to system dynamics: An autobiography, in management laureates. JAI Press
11. Ford A, Cavana R (eds) (2004) Special Issue of the Syst Dyn Rev
12. Hardin G (1968) The tragedy of the commons. Science 162:1243–1248
13. IPCC (1997) An introduction to simple climate models used in the IPCC second assessment report. ISBN 92-9169-101-1
14. IPCC (2007) Climate change 2007: The physical science basis, summary for policymakers. www.ipcc.ch/
15. Kump L (2002) Reducing uncertainty about carbon dioxide as a climate driver. Nature 419:188–190
16. Meadows DH, Meadows DL, Randers J, Behrens W (1972) The limits to growth. Universe Books
17. Morecroft J (2007) Strategic modelling and business dynamics. Wiley, Chichester
18. Richardson J, Pugh A (1981) Introduction to system dynamics modeling with dynamo. Pegasus Communications
19. Sterman J (2000) Business dynamics. McGraw-Hill, Irwin
20. Sterman J (ed) (2002) Special Issue of the Syst Dyn Rev
21. Sterman J, Sweeney L (2007) Understanding public complacency about climate change. Clim Chang 80(3–4):213–238
22. Warren K (2002) Competitive strategy dynamics. Wiley, Chichester
23. Webster M et al (2003) Uncertainty analysis of climate change and policy response. Climat Chang 61:295–320

Books and Review

Houghton J (2004) Global warming: The complete briefing, 3rd edn. Cambridge University Press, Cambridge

System Dynamics Models, Optimization of

BRIAN DANGERFIELD
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

Article Outline

Glossary

Econometrics A statistical approach to economic modeling in which all the parameters in the structural equations are estimated according to a 'best fit' to historical data.

Maximum likelihood A statistical concept which underpins calibration optimization and which generates the most likely parameter values; it is equivalent to the parameter set which minimizes the chi-square value.

Objective function See **Payoff** below.

Optimization The process of improving a model's results in terms of either an aspect of its performance or by calibrating it to fit reported time series data.

Payoff A formula which expresses the objective, say, maximization of profits, minimization of costs or minimization of the differences between a model variable and historical data on that variable.

Zero-one parameter A parameter which is used as a multiplier in a policy equation and serves the effect of bringing in or removing a particular influence in determining the optimal policy.

Definition of the Subject

The term 'optimization' when related to system dynamics (SD) models has a special significance. It relates to the mechanism used to improve the model vis-à-vis a criterion. This collapses into two fundamentally different intentions. Firstly one may wish to improve the model in terms of its performance. For instance, it may be desired to minimize overall costs of inventory whilst still offering a satisfactory level of service to the downstream customer. So the criterion here is cost, and this would be minimized after searching the parameter space related to service level. The direction of need may be reversed and maximization may be desired as, for instance, if one had a model of a firm and wished to maximize profit subject to an acceptable level of payroll and advertising costs. Here the parameter space being explored would involve both payroll and advertising parameters. This type of optimization might be described generically as *policy optimization*.

Optimization of performance is also the *raison d'etre* of other management science tools, most notably mathematical programming. But such tools are usually static: they offer the 'optimum' resource allocation given a set of constraints and a performance function to either maximize or minimize. These models normally relate to a single time point and may then need to be re-run on a weekly or monthly basis to determine a new optimal resource allocation. In addition, these models are often linear (certainly so in the case of linear programming), whereas SD models are usually non-linear. So the essential differences are that SD model optimization for performance involves both a dynamic and a non-linear model.

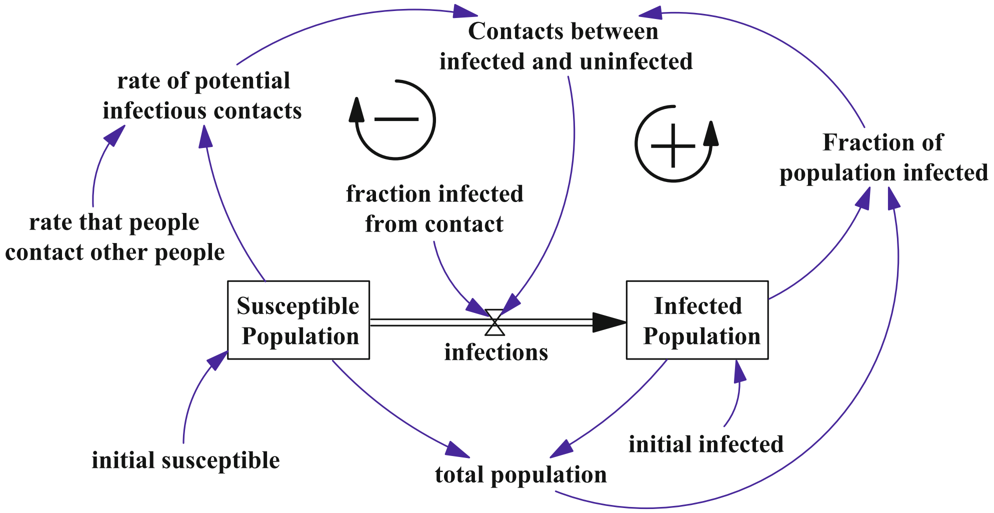
A separate improvement to the model may be sought where it is required to fit the model to past time series data. Optimization here involves minimizing a statistical function which expresses how well the model fits a time-series of data pertaining to an important model variable. In other words a vector of parameters are explored with a view to determining the particular parameter combination which offers the best fit between the chosen important model variable and a past time series data set of this variable. This type of optimization might be generically termed *model calibration*. If *all* the parameters in the SD model are determined in this fashion then the process is equivalent to the technique of econometric modeling. A good comparison between system dynamics and econometric modeling can be found in Meadows and Robinson [12].

Optimization as Calibration

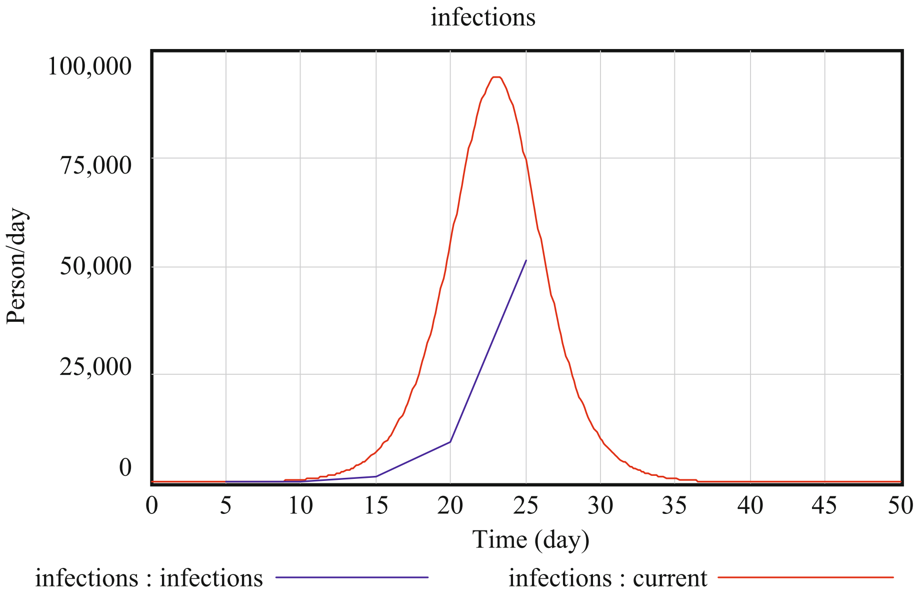
In these circumstances we wish to determine optimal parameters, those which, following a search of the parameter space, offer the best fit of a particular model variable to a time series dataset on that variable taken from real world reporting.

As an example consider a variation of the one of the epidemic models which are made available with the Vensim™ software. The stock-flow diagram is presented as Fig. 1.

In this epidemiological system members of a susceptible population become infected and join the infected population. Epidemiologists call this an S–I model. It is a sim-


Contacts between infected and uninfected
rate of potential infectious contacts
Fraction of population infected
fraction infected from contact
rate that people contact other people
Susceptible Population
infections
Infected Population
initial susceptible
total population
initial infected

System Dynamics Models, Optimization of, Figure 1
Stock-flow diagram for a simple epidemic model



System Dynamics Models, Optimization of, Figure 2
Current (base) run of the model and reported data on infections

pler variation of the S–I–R model which includes recovered (R) individuals.

Suppose some data on new infections (at intervals of five days) are available covering 25 days of a real-world epidemic. The model is set with a time horizon of 50 days which is consistent with, say, a flu epidemic or an infectious outbreak of dysentery in a closed population such as a cruise ship. The 'current' run of the model is shown in Fig. 2, with the real-world data included for comparison.

Clearly there is not a very good correspondence between the actual data and the model variable for the infection rate (infections). We wish to achieve a better calibration, and so there is a need to select relevant parameters through which the calibration optimization can be performed over. Referring back to Fig. 1, we can see that the *fraction infected from contact* and the *rate that people contact other people* are two possible parameters to consider. The initial infected and initial susceptible are also parameters of the model in the strict sense of the term, but we will ignore them on this occasion. In this model the *initial infected* is 10 persons and *initial susceptibles* number 750,000 persons.

The chosen value for the *fraction infected from contact* is 0.1, while that for the *rate that people contact other people* is 5.0. The former is a dimensionless number while the latter is measured as a fraction per day (1/day). This is obtained from consideration of the *rate of potential infectious contacts* (persons/day) as a proportion of the *susceptible population* (persons).

The optimization process for calibration involves reading into the model the time series data, in this case on new infections, and, secondly, determining the range for the search in parameter space. There is usually some basic background knowledge which allows a sensible range to be entered. For instance, a probability can only be specified between 0 and 1.0. In this case we have chosen to specify the ranges as follows:

$$0.03 \leq \text{fraction infected from contact} \leq 0.7$$

$$2 \leq \text{rate that people contact other people} \leq 10\,.$$

A word of warning is necessary in respect of optimizing delay parameters. Because there is a risk of mathematical instability in the model if the value of DT (the TIME STEP) is too large relative to the smallest first-order delay constant, it is important to ensure the TIME STEP employed in the model is sufficiently small to cope with delay constant values which may be reached during the search of the delay parameter space. In other words ensure the minimum number for the search range on the delay parameter is at least double the value of the TIME STEP.

Maximum Likelihood Estimation and the Payoff Function

The optimization process involves a determination of what are termed statistically as maximum likelihood estimates. In Vensim™ this is achieved by maximizing a payoff function. Initially this is negative and the optimization process should ensure this becomes less negative. An ideal payoff value, after optimization, would be zero. A weighting is needed in the payoff function too, but for calibration optimization this is normally 1.0. Driving the payoff value to be larger by making it less negative has parallels with the operation with the simplex algorithm common in linear programming. This algorithm was conceived initially for problems where the objective function was to be minimized. Its use on maximization problems is achieved by minimizing the negative of the objective function.

System Dynamics Models, Optimization of, Table 1
Data used for calibration experiment

Time	5	10	15	20	25
Infections	30	230	1400	9500	51 400

During the calibration search, Vensim™ takes the difference between the model variable and the data value, multiplies it by the weight, squares it and adds it to the error sum. This error sum is minimized. Usually data points will not exist at every time point in the model. Here the model TIME STEP is 0.125 (1/8th), but let us assume that reported data on new infections have been made available only at times $t = 5, 10, 15, 20$ and 25 so the sum of squares operation is performed only at these five time points.

The data are shown as Table 1.

The Recording Point for Reported Data

System dynamics models differentiate between stock and flow variables and the software used for simulating such models advances by a small constant TIME STEP (also known as DT). This has implications for the task of fitting real-world reported data to each type of system dynamics model variable. The following is the issue: at what point in a continuum of time steps should the reported data be recorded at? This is important because the reported data has to be read into the model to be compared with the simulated data. The answer will be different for stock and flow variables.

Where the reported data relate to a stock variable the appropriate time point for recording will be known. If it is recorded at the end of the day (say a closing bank balance) then the appropriate point for data entry in the model will be the beginning of the next day. Thus the first data point above is at time $t = 5$ (5.00) and would, if it were a stock, correspond to a record taken at the very end of time period 4.

However, if the data relate to a flow variable, as in the case of new infections here, the number is the total new infections which have occurred over the entire time unit (day, week, month etc.) and so there is a decision to be reached as to which time point the data are entered at. This is because the TIME STEP (DT) is hardly ever as large as the basic time unit which the model is calibrated in. The use of 5 (10, 15 etc.) above implies that the data on new infections over the period of time $t = 0$ to $t = 5$ are compared with the corresponding model variable at time 5 + 1 ∗ DT (and the new infections over the period $t = 5$ to $t = 10$ at time 10 + 1 ∗ DT etc.). A more appropriate selection might be towards the end of the 5-day time period. Following the example above using a TIME STEP = 0.125, this might be at time 4 + 7 ∗ DT (that is at 4.875).

Calibration Optimization Results

Based upon the data on new infections shown above and the chosen ranges for the parameter search, the following output is obtained (Table 2). After 114 simulations the optimized values for our two parameters are shown to be 0.08 and 5.12 and the payoff is over 2500 times larger (less negative). Replacing the original parameters with the optimized values reveals the result shown in Fig. 3. To take things further we may wish to put confidence intervals on the estimated parameters. One way of accomplishing this is by profiling the likelihood and is described in Dangerfield and Roberts [3].

Avoid Cumulated Data

There might be a temptation to optimize parameters against cumulated data when the data are reported essentially as a flow, as is the case here. Were the data to be cumulated we would obtain as shown in Table 3.

The results from this optimization are shown in Table 4. The ranges for the parameter space search are kept the same but the payoff function now involves a comparison of the model variable *infected population* with the corresponding cumulated data. Figure 4 shows the resultant fit to infected population is good, but that is manifestly not borne out when we consider the plot of infections obtained from the same optimization run (Fig. 5).

System Dynamics Models, Optimization of, Table 2
Results from the calibration optimization

Initial point of search
fraction infected from contact = 0.1
rate that people contact other people = 5
Simulations = 1
Pass = 0
Payoff = −2.67655e + 009

Maximum payoff found at:
fraction infected from contact = 0.0794332
*rate that people contact other people = 5.11568
Simulations = 114
Pass = 6
Payoff = −1.06161e + 006

System Dynamics Models, Optimization of, Table 3
Cumulated reported data for the infected population

Time	5	10	15	20	25
Infected population	30	260	1660	11 160	62 560

System Dynamics Models, Optimization of, Table 4
Results from the calibration using cumulated data

Initial point of search
fraction infected from contact = 0.1
rate that people contact other people = 5
Simulations = 1
Pass = 0
Payoff = −2.48206e + 011

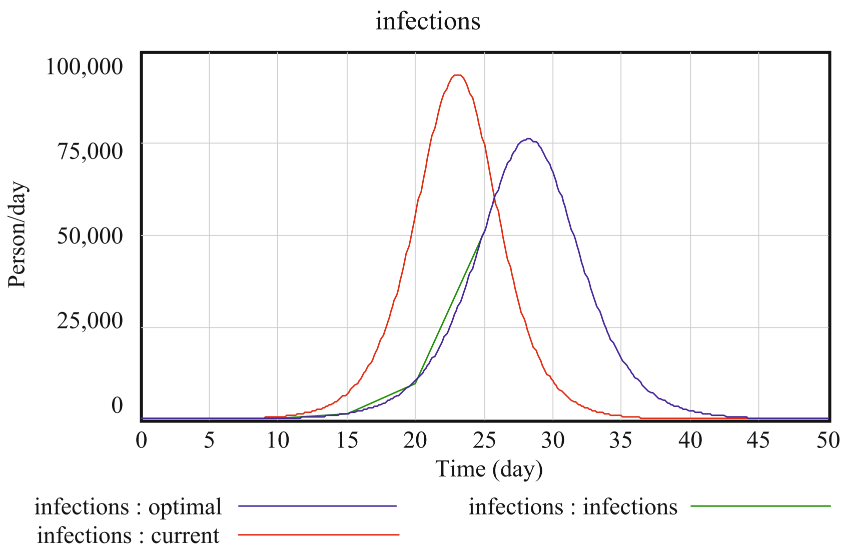
Maximum payoff found at: fraction infected from contact = 0.0726811 *rate that people contact other people = 4.96546 Simulations = 145 Pass = 6 Payoff = −212 645 The final payoff is −212 645

The reason for this is rooted in statistics. The maximum likelihood estimator is equivalent to the chi-squared statistic. This is turn assumes that each expected data value is independent. A cumulated data series would not exhibit this property of independence.

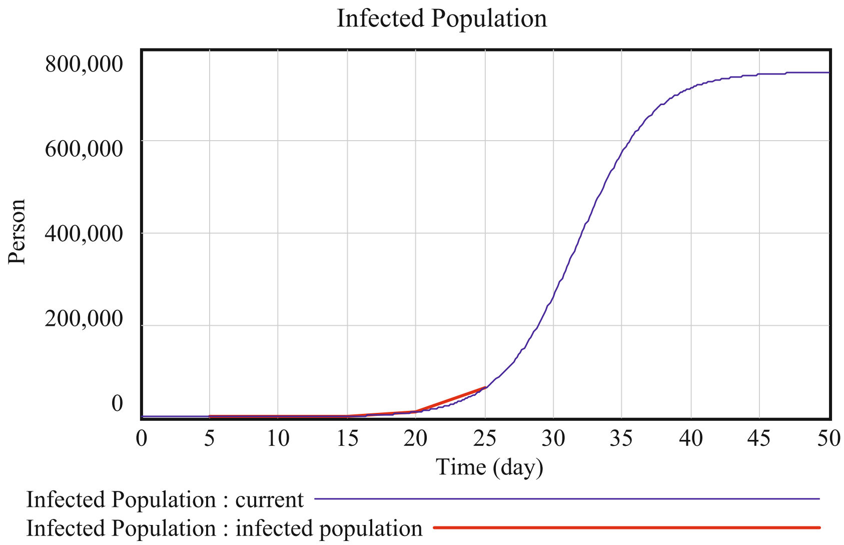
As an aside it is worth pointing out that this model, with suitable changes to the variable names and the time constants involved, could equally represent the diffusion of a new product into a virgin market. In systems terms the structures are equivalent. The *fraction infected from contact* is the same as, say, the fraction reached by word of mouth or advertising and the *rate that people contact other people* is a measure of the potential interactions at which new products might be mentioned amongst the members of the relevant market segment. An *infected population* is equivalent to a customer base, the number of adopters of the relevant product. So it is possible to shed light on important real-world marketing parameters through a calibration optimization of models of this general structure.

Optimization of Performance (Policy Optimization)

An example model is to be used to illustrate the process of optimization to improve the performance of the system, and this is illustrated in Fig. 6. It concerns the service requirements which can arise following the sale of a durable good. These items are typically sold with a 12-month warranty and during this time the vendor is obliged to offer service if a customer calls for it. In this particular case the vendor is not being responsive in terms of staffing the service section. The result is that as sales grow the increasing number of service requests is putting pressure on the



System Dynamics Models, Optimization of, Figure 3
Reported data on infections and optimized (calibrated) model; the base case (current) is reproduced for reference

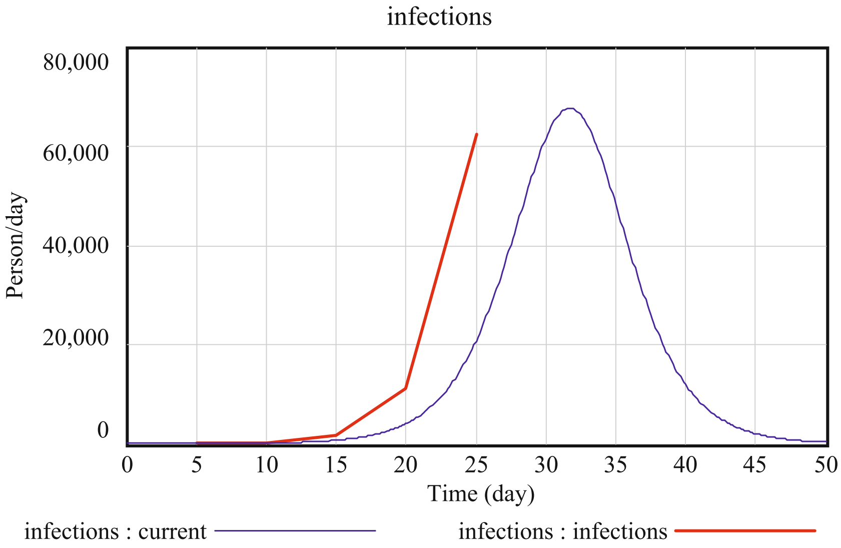


System Dynamics Models, Optimization of, Figure 4
The cumulative model variable (infected population) together with reported data

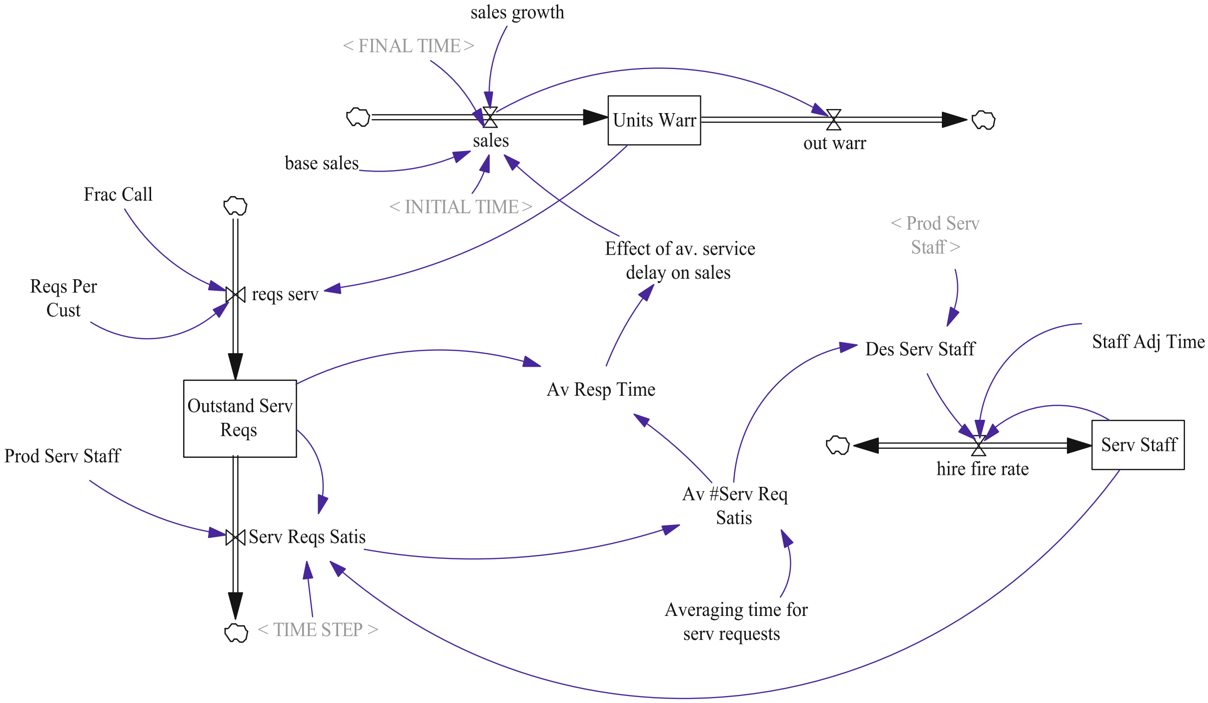
service personnel. The delay in responding to service calls also increases and the effect of this is that future sales are depressed because of the vendor's acquired reputation for poor service response. The basic behavior mode is overshoot and collapse.

In the model depicted in Fig. 6, the growth process is achieved by a RAMP function which causes sales of the good to increase linearly by 20 units per month from a base of 500 units per month.

The payoff function is restricted to the variable *Sales*. However, this need not be the case. Where a number of variables might be options in a payoff function, it is possible to assign weights to each such that the sum of the weights is 1.0 (or 100). The optimization process will then proceed with the software accumulating a weighted payoff which it will attempt to maximize. Weights are positive when more is better and negative when less is better.



System Dynamics Models, Optimization of, Figure 5
The corresponding fit to infections is poor



System Dynamics Models, Optimization of, Figure 6
Model of service delays for durable goods under warranty

Policy Experiment No. 1

Here it is decided to try to improve the productivity of the service staff. Currently they manage, on average, to respond to 120 calls per operative per month. It may be an option to improve their productivity by, say, providing them with hand-held devices which direct each operative from one call to the next – calls which may have arisen since setting out from their base. In this way their call routing is improved.

System Dynamics Models, Optimization of, Table 5
Optimization results for the productivity of the service staff

Initial point of search
Prod Serv Staff = 120
Simulations = 1
Pass = 0
Payoff = 27 743.5

Maximum payoff found at:
*Prod Serv Staff = 122.647
Simulations = 27
Pass = 3
Payoff = 29 915

The optimization parameter is *Prod Serv Staff*, and we select an upper limit for the search range of 240 calls per person per month. The chosen performance variable is *Sales*, since we wish to maximize this – or at least not have it overly depressed by poor response times. The results are shown in Table 5. We see that the payoff is increased and that the optimum productivity is a modest increase of 2.6 requests per month, on average. This should be easily achievable and perhaps without expenditure on high-tech devices. The graphical output for sales is shown in Fig. 7.

For comparison, the effect of increasing the productivity to as high as 150 calls per month, on average, is also shown. This would represent an increase of 25% and would be much more difficult to accomplish. Here the benefit of optimization is highlighted. A modest increase in productivity returns a visibly improved sales performance (although the basic behavior mode is unchanged), whilst a much greater productivity increase offers little extra benefit for the effort and cost involved in improving productivity.

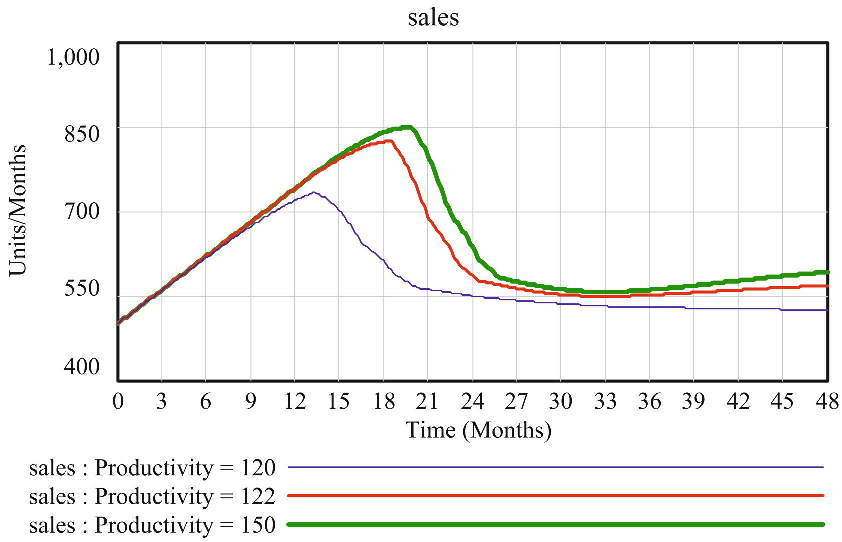
Policy Experiment No. 2

Another approach to policy optimization involves the use of a zero-one parameter which has the effect of either including or excluding an influence on policy. Suppose it was thought that the quantity of product units in warranty should exert an influence on the numbers of service personnel hired (or fired). The equation for the desired number of service staff (*Des Serv Staff*) can be expressed as:

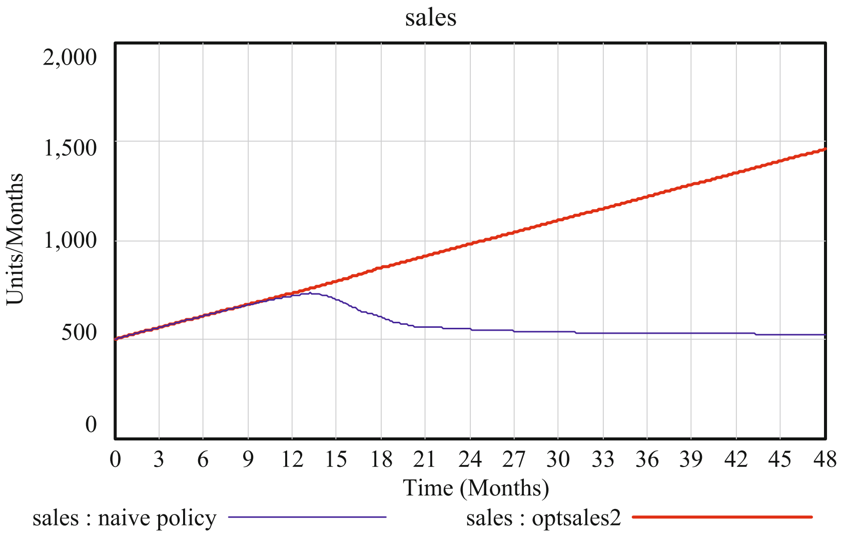
Des Serv Staff = "Av #Serv Req Satis"/Prod Serv Staff ∗ trigger + ("Av #Serv Req Satis"/Prod Serv Staff)∗(Units Warr/initial units in warranty)∗(1-trigger). (Units: Persons).

The *trigger* variable is initially set to 1.0 and so the more sophisticated policy is not active. The optimization run results are shown in Table 6. Clearly there is benefit from including the more sophisticated policy which takes into account the current numbers of product units in warranty.

The graphical output is unequivocal (Fig. 8). Sales are continuously increasing when the recruitment policy for service personnel takes into account the number of product units in warranty. The depressive effect on sales of poor service performance is non-existent.



System Dynamics Models, Optimization of, Figure 7
Plots of sales achieved for differing productivities



System Dynamics Models, Optimization of, Figure 8
Comparison of sales from two different policy drivers

System Dynamics Models, Optimization of, Table 6
Optimization results from selection of policy drivers

Initial point of search
trigger = 1
Simulations = 1
Pass = 0
Payoff = 27 743.5

Maximum payoff found at:
*trigger = 0
Simulations = 13
Pass = 3
Payoff = 47 090.9

Whilst this might seem an obvious policy, it is surprising how easily the naïve alternative might be accepted without question. The number of calls a typical operative can manage each month is well known along with the (historical) number of service requests satisfied. Hence, the desired number of staff is more or less fixed. This comes undone when there is a growth in the number of products sold. In this different environment such a simplistic policy can, as shown, lead to overshoot and collapse. Notice needs to be taken of the changing number of product units in warranty in order that a more effective system performance is achieved.

The above experiments are illustrative only, and there is no intention of over-working a simple teaching model in order to uncover an ideal policy. In the case of policy optimization a wide range of possible alternatives exists. Indeed, a process of learning naturally arises through carrying out repeated optimization experiments with the model [1].

Examples of SD Optimization Reported in the Literature

Amongst the earliest work in this area the writings of Keloharju are worthy of mention. He contributed a number of papers on the topic in the pages of *Dynamica*. See, for example, Keloharju [9]. His work brought the concept to prominence but he did not employ the method on anything other than problems described in text books or postulated by himself. For instance, an application of optimization to the project model contained in Richardson and Pugh's [13] text is contained in Keloharju and Wolstenholme [11]. A statement of the method together with some textbook examples is also available [10]. Additionally, an overview of the methods and their deployment on textbook examples has been contributed by the current author [3]. Finally, there is an example of optimization applied to defence analysis. Again though it is a standard defence model – the armored advance model – rather than any real-world study [14].

Retaining the emphasis on textbook problems for the moment, Duggan [6] employs Coyle's model [1] of the *Domestic Manufacturing Company* to illustrate the methods of multi-objective optimization – an advance over standard SD optimization with its single objective function. The concept of multiple objectives arises from multi-criteria decision-making where a situation can be judged on more than one performance metric. While a multi-objective payoff function can be formulated using a set of weights, it is argued that the selection of the weights is very individual-specific. The multi-objective approach – underpinned by the methods of genetic algorithms – rests upon determining a Pareto-optimal situation, defined as one where no improvement is possible without making some other aspect worse. In other words the method strives for an optimal solution which is not dominated by any other solution. The author demonstrates the approach combining two objectives in the model: one for the differences between desired stock and actual and another between desired backlog and actual.

In terms of applications to real-world problems, the current author has also used the methods of optimization in research conducted in connection with modeling the epidemiology of HIV/AIDS. Fitting a model of AIDS spread to data was carried out for a number of European countries [2,4]. The optimized parameters furnished support for some of the features of AIDS epidemiology which, at the time, were being uncovered by other branches of science. For example, the optimized output revealed that a U-shaped profile of infectiousness in a host was necessary in order to achieve a best fit to data on new AIDS cases. This infectiousness profile was also evidenced by virologists who had analyzed patients' blood and other secretions on a longitudinal basis.

Within this strand of research, a much more complex optimization was performed using American data on transfusion-associated AIDS cases [5]. The purpose here was to estimate the parameters for a number of plausible statistical HIV incubation distributions. Given the nature of the data, the point of infection could be quite accurately determined, but two difficulties were evident: the data were right-censored and the number receiving infected transfusions in each quarter was unknown. However, the SD optimization could estimate this number as part of the process, in addition to estimating parameters of the incubation distribution. The best fit was found to be a three stage distribution similar to the gamma and one which accorded with the high-low-high U-shaped infectiousness profile which was receiving support from a number of sources.

In the marketing domain Graham and Ariza [8] carried out an optimization on a system dynamics model which was designed to shed light on the allocations to make from a marketing budget in a high-tech client firm. Assuming the budget was fixed, the task was to optimize the allocations across more than 90 'buckets' – combinations of product lines, marketing channels and types of marketing. However, these were not discrete: advertising on one product line might have crossover effects on another and the impacts could propagate over a period of time. One major conclusion for this firm was that the advertising allocation should be increased markedly. In general intuitive allocations were shown to fall short of the ideal: they were directionally correct but magnitudes fell short often by factors of three or four.

Future Directions

A primary aim must be to see more published work which describes optimization studies carried out on real-world SD applications. There may be frequent use of optimization in consulting assignments but such activities are rarely published. The references herein suggest that, thus far, outside of unpublished work, the number may be three at most. Whilst software requirements may have inhibited use of SD optimization in the past, there are now no computational barriers to its use and it is to be hoped that in future this quite powerful analytical tool in SD will feature in more application studies.

An advance in the methodology itself has been developed by Duggan [7] and this is a promising pointer for the future. Based on genetic algorithms, it is best suited to the class of SD problems that are agent-based and this highlights a slight limitation. Traditional optimization takes the policy equations as given and explores the parameter space to determine an optimal policy. Instead he has offered an approach which searches over both parameter space and policy (strategy). Theoretically there is no limit to the number of strategies which can be evaluated in this approach, although the user has to define a set in advance of the runs. Under a conventional optimization approach a limited tilt at this is possible using the zero-one parameter method suggested above, although this would restrict the enumerated strategies to two only. Duggan demonstrates the new approach using a classic SD problem: the four agent beer-game. We await its use in a real-world application.

Bibliography

1. Coyle RG (1996) System dynamics modelling: a practical approach. Chapman & Hall, London
2. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five European countries. In: O.R. Work in HIV/AIDS, 2nd edn. Operational Research Society, Birmingham, pp 7–13

3. Dangerfield BC, Roberts CA (1996) An overview of strategy and tactics in system dynamics optimisation. J Oper Res Soc 47(3):405–423
4. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
5. Dangerfield BC, Roberts CA (1999) Optimisation as a statistical estimation tool: an example in estimating the AIDS treatment-free incubation period distribution. Syst Dyn Rev 15(3):273–291
6. Duggan J (2005) Using multiple objective optimisation to generate policy insights for system dynamics models. In: Proceedings of the international system dynamics conference, Boston. System Dynamics Society. (CD-ROM)
7. Duggan J (2008) Equation-based policy optimisation for agent-oriented system dynamics models. Syst Dyn Rev 24(1):97–118
8. Graham AK, Ariza CA (2003) Dynamic, hard and strategic questions: using optimisation to answer a marketing resource allocation question. Syst Dyn Rev 19(1)27–46
9. Keloharju R (1977) Multi-objective decision models in system dynamics. Dynamica 3(1)3–13 and 3(2)45–55
10. Keloharju R, Wolstenholme EF (1988) The basic concepts of system dynamics optimisation. Syst Pract 1:65–86
11. Keloharju R, Wolstenholme EF (1989) A case study in system dynamics optimisation. J Oper Res Soc 40(3):221–230
12. Meadows DM, Robinson JM (1985) The electronic oracle. Wiley, Chichester (Now available from the System Dynamics Society, Albany NY)
13. Richardson GP, Pugh AL (1981) An introduction to system dynamics modelling with DYNAMO. MIT Press, Cambridge (Now available from Pegasus Communications, Waltham, MA)
14. Wolstenholme EF, Al-Alusi AS (1987) System dynamics and heuristic optimisation in defence analysis, Syst Dyn Rev 3(2):102–115

System Dynamics Philosophical Background and Underpinnings

CAMILO OLAYA
Universidad de Los Andes, Bogotá, Colombia

Article Outline

Glossary

Philosophy The reflection and study of our most basic assumptions – or the assumptions themselves.

Mental model A mental image of selected concepts and relationships of the world around us which we consider relevant for explaining the behavior of a particular system.

Presentationalism Synonymous of idealism. The view that material objects or external realities do not exist apart from our knowledge or consciousness of them.

Definition of the Subject

We all tend to take things for granted. Indeed it is a common place to judge formal models exclusively based on the technical grounds and on the logic with which those models were built without a proper reflection on the assumptions underlying those models. This omission is even more pressing in complexity and system science, since these areas represent a novel challenge for philosophers of science – e. g. see an overview in [34].

What is the idea of reality with which we work? What do we assume about human nature? What kind of knowledge do we pursue? What kind of knowledge do we obtain? What is the scope of rational inquiry? What are the basis and the implications of our own reasoning methods? The identification of how philosophy has shaped the work of scientists – on a conscious or unconscious level – is essential for comprehending the implications, the limitations, and the scope of our very scientific practice. The lack of concern by scientists for these issues may explain many of their failures which has produced just a sort of inertial blindness that is easy to recognize in current scientific debates.

One of the strengths of system dynamics (abbreviated, SD) is that it leads us to make explicit our assumptions about the systems we deal with. This attitude, i. e. the importance of reflection upon our own assumptions, is also fundamental for the very development and practice of system dynamics. Many of the debates on different issues of every day scientific practice such as model conceptualization, formal model building, validation, policy design, etc. are informed and can be enlightened by the reflection on the philosophical background behind those processes. There are also various fundamental aspects of SD that are yet to be demarcated, e. g. the characterization of SD explanations. Furthermore, the ambiguity of the discussions found in large part of related literature, characterized by superficiality, confusion of terms, misdirected arguments, etc. only adds noise and it complicates the advance of a discipline. This article sketches and overview on some basic assumptions regarding the development and the practice of system dynamics. Various suggestions that help to integrate various debates are introduced and important clarifications are also indicated.

Introduction

The philosophical background and underpinnings of a discipline should have to do with its most basic universal assumptions. Such a discussion becomes difficult if we bear in mind that those assumptions are not necessarily shared by practitioners and researchers. Nevertheless, central premises can be identified which in turn can be related to important questions regarding philosophical concerns such as reality and knowledge.

The article is organized as follows. After this short introduction, the second sections develops an overview of the origins of system dynamics underlining fundamental aspects that formed what can be called the core of the discipline. This historical review highlights the initial interest, purposes and initial assumptions around the foundation of SD. With these elements the following sections introduce various philosophical issues that can be identified underlying system dynamics. Perhaps the central aspect is presentationalism, a stance associated with the notion of "mental models" which is central in SD; this is the topic of the fourth section. The following section makes a clarification on the controversial issue of positivism and relates presentationalism with knowledge. The sixth section summarizes the position of system dynamics regarding social theory. The seventh section presents the inquiry of explanation clarifying that in spite that SD involves causal models the nature of its kind of explanation can be found in the notion of mechanism. The eighth section introduces the implication of the use of computer simulation as a distinctive epistemology which is different from the traditional discourse in philosophy of science. The ninth section outlines future directions.

Before starting, a brief warning should be made: given the scope of this review and the limited space for covering very wide subjects then this article should be viewed as a broad introductory overview of the different topics. The cited literature has been selected as possible starting points for further inquiry.

System Dynamics

In order to address a "philosophical background" the first question naturally would be: What is system dynamics? Already this inquiry can be a matter of debate, e. g. [102]. Indeed SD has been labeled as a theory [23,49], a method [18,56,63,98,108], a methodology [81], a field of study [17,78], a tool [61], a paradigm, among other nouns. A natural starting point is the work of Jay Forrester, the founder of system dynamics. A brief historical review should help to grasp the very core we are looking for since it can show the initial motivations, assumptions, and purposes behind the development of SD.

Genesis of System Dynamics

Jay W. Forrester, member of the Sloan School of Management at MIT, was looking for linkages between engineering and management education given his background in feedback control systems and computers [28]. In 1956 he wrote a "note" to the Faculty Research Seminar, the first ever MIT "D-memo"; in this communication he sketched the worldview of what would be known as "system dynamics" [32].

He started with a strong criticism of economic models. The following were the central aspects: (i) their failure to reflect adequately the loop structures that make up economic systems; in particular this neglect leads to exclude inherent properties of closed loops such as resistance to change, accumulations and delays; (ii) The incapacity for including flows of goods, money, information, and labor, in one single interrelated model; (iii) The exclusion of changing mental attitudes that affect and explain economic processes; (iv) The use of linear equations for describing systems; (v) The restriction of building models constrained by the capacity for manipulating numerical data and solving the equations; (vi) The overconfidence in multiple regression analysis for obtaining coefficients for equations that define economic behavior; (vii) The lack of reflection and discussion on the very assumptions underlying every model preferring an emphasis on the logic with which the model is developed.

After delineating these points, Forrester then proceeded in the same note to highlight techniques that were largely underused at that time: servomechanisms, differential equations, and what he called "the art of simulation". Anchored on the mentioned assessment and on these developments Forrester conceived "a new avenue of attack for understanding the firm and the economy" ([28] p. 336) envisaging *a new kind of models* that would include aspects such as:

Dynamic structure: Detailed attention to the *sequences of actions* which occur in the system being studied and to the *forces* which trigger or temper such actions, with a particular concern on the controlling influences of lags and delays.

Information flows: Explicit recognition of information flow channels and information transformation with time and transmission.

Decision criteria: Re-examination of the proper decision criteria which must not be defined as depending only on current values of gross economic variables; instead, such criteria must be traced to the motivations, hopes, objectives and optimism of the people involved, including as well what he calls business man's intuition which "represents a disordered accumulation of basic insights into how people and social systems react. The hope for the future lies in generating an orderly arrangement of basic insights" ([28] p. 342).

Non-linear systems: Economic systems present most – if not all – of the time highly non-linear characteristics.

Differential equations: The behavior of economic systems should be better described by non-linear differential equations since they have been developed to describe delays, momentum, elasticity, reservoirs, and accelerations, which are better suited quantities for describing the economic world. In practice these equations would be handled as incremental difference equations in order to obtain numerical solutions.

Incremental changes in variables: To prefer the formulation of a model in terms of the motivations that cause incremental changes in a variable since the new value of a variable "can be found by solving the equations for its incremental change and then adding the change to the preceding full value of the variable" ([28] p. 344).

Model complexity: Much complex and complete models can be developed with these techniques.

Empirical solutions: It is useless to look for explicit unique or "correct" solutions; instead, these models provide diverse solutions according to the different assumptions about the model structure and the initial values of the variables.

Symbolism and correspondence with real counterparts: The possibility of having a pictorial representation – a flow diagram – whose processes of information, money, goods, and people, are moved, i. e. simulated, time-step-by-time-step from place to place.

Structure over coefficient accuracy: To prefer a structure in which we have confidence using intuitively estimated coefficients instead of using unlikely structures with accurately derived coefficients from statistical data.

A subsequent advance came in 1958 with an article entitled "Industrial Dynamics: A Major Breakthrough for Decision Makers" published in the *Harvard Business Review* [27]. In this article Forrester shaped the previous ideas with the concern that management should evolve from a highly fragmentized art to a profession capable of recognizing unified systems given that the task of management is to interrelate the flows of information, materials, labor, money, and capital equipment. He again emphasized features such as electronic data processing, decision making, simulation, feedback control, and information flows. These elements were presented as the cornerstones of the innovative industrial dynamics program at MIT.

Industrial Dynamics

The definitive breakthrough came in 1961 with his *magnus opus* "Industrial Dynamics" [24]. The main motivation behind was the development of a *science* for designing and controlling industrial systems, the quest for a management science. In particular he conceived it as "a method of system analysis for management." (p. 9). He stated:

> Industrial dynamics is the study of the information-feedback characteristics of industrial activity to show how organizational structure, amplification (in policies), and time delays (in decisions and actions) interact to influence the success of the enterprise (p. 13).

Forrester underlined four pillars for this new science: information-feedback control theory anchored on the concept of servomechanisms, the study of decision-making processes, an experimental approach to system analysis based on simulation, and the use of computers.

Coming from engineering, he had in mind models that deal with nonlinear dynamic systems whose purpose is to *design* new systems – as opposed to just *explain* systems; the models should show how changes in policies or structure will produce better or worse behavior. In order to accomplish this aim Forrester indicated that we should focus on understanding the characteristics of the system in hands (instead of looking for specific predictions) and on our assumptions about them. "We then have a means for tracing the implications of our assumptions" ([24] p. 55).

What should be included in a model? Forrester underlined that "there will be no such thing as *the* model of a social system, any more than there is *the* model of an aircraft … the factors that must be included arise directly from the questions that are to be answered" ([24] p. 60). It is expected that these factors will include closed-loop information-feedback structures that give rise to so much of the interesting behavior. An important aspect of this new kind of models is symbolism and pictorial representations by means of flow diagrams with a special emphasis on correspondence: the model variables should correspond to those in the system being represented. In this book Forrester also demarcated the network structure of this new kind of models as made of four basic components: accumulations (levels), flows (rates), decision functions, and information channels; these networks trace cause-effect relationships which are described via mathematical formalization. He also discussed in detail how to represent delays and how to model decision processes which are particularly defined by general policies, i. e. rules that state how operation decisions are made converting information into action; this study of general policies explains the importance for SD of the examination of human decision-making processes. Another fundamental characteristics are continuous flows and aggregation: "grouping of individual events into classes … Our interest in the model … is from the viewpoint above the separate individual transactions" ([24] p. 65); it is assumed that different individual items are controlled by the same identical decision-function; this notion leads to aggregation which is a distinctive aspect of SD: "items controlled by sufficiently similar policies that depend on sufficiently equivalent information sources may be combined into a single channel" ([24] p. 109); the central criterion for such aggregation is the purpose of the model. Finally, model significance (or validity) rests on its suitability for a particular purpose which is motivated by the design of improved industrial and economic systems.

Principles of System Dynamics

The main concern in these initial steps were business operations. However, Forrester sketched a glance to a broader view in the final part of his book; there, he speaks of "system" dynamics since "the study of systems can provide a framework to unite subjects … The dynamic model represents a system as broad as one chooses to describe" ([24] pp. 344, 346). Indeed, looking for a more general view he presents a section of "principles of systems structure":

> The principles to be discussed here all arise in the context of information-feedback systems. They are systems principles. They are not the principles of the management art such as have been taught in organization, production, and human relations courses. Because the principles apply to systems behavior, they do not fall into neat separate packages … The concept of a system implies interaction and interdependence. In attempting to identify factors that are common to all systems, we must keep the essential indivisibility in mind ([24] pp. 347, 348).

Indeed, Forrester reaffirmed this general *systems* view in a follow-up book entitled *Principles of Systems* [25] published in 1968. He gives a basic definition of system as "a grouping of parts that operate together for a common purpose" ([25] p. 1-1). Moreover, he suggests that the way to organize knowledge is with this idea of system which are represented by the models we develop:

> A structure (theory) is essential if we are to effectively interrelate and interpret our observations in any field of knowledge … Without an organizing structure, knowledge is a mere collection of observations, practices and conflicting incidents … A model is s substitute for an object or system … Any set of rules and relationships that describe something is a model of that thing. In this sense, all of our thinking depends on models. Our mental processes use concepts which we manipulate into new arrangements. These concepts are not, in fact, the real system that they represent. The mental concepts are abstractions based on our experience. This experience has been filtered and modified by our individual perception and organization processes to produce our mental models that represent the world around us ([25] pp. 1-2, 3-1).

The previous statement summarizes core assumptions for SD. It points at the central notion of *mental model* and it emphasizes that these models are models of *systems*. In fact, in an article published in 1968 in *Management Science*, Forrester [28] underlined that this application of feedback concepts to social systems was evolving toward a theory of structure in systems with a particular goal of policy design. Specifically, "industrial dynamics is a philosophy of structure in systems. It is also gradually becoming a body of principles that relate structure to behavior" ([28] p. 141). Two fundamental variables are identified: levels and rates, and the basic structural element is the feedback loop: "every decision is responsive to the existing condition of the system and influences that condition" ([28] p. 143). And, as stated above, the structure is an aid to organize knowledge in a particular situation given a particular purpose which is motivated by the pursue of explanation and policy design.

This historical review has accentuated central aspects of the foundation of SD. As a summary, this science initially envisaged by Forrester for designing systems is known nowadays as system dynamics. The models developed in SD have distinctive characteristics: dynamic structures, information flows, study of decision criteria, non-linearity, difference equations, symbolism and correspondence, and emphasis on confidence based on the structure of the model. The practice of this science is anchored on: the concept of servomechanism, the study of decision-making processes, the embrace of an experimental approach, and the use of computers for building formal models. These models are models of systems and the main goal is to help to organize our knowledge – which can be seen as arranged in mental models – so as to enhance learning processes and systems design on concrete settings and under specific purposes.

With these elements in mind, it is now possible to proceed with a brief discussion on important aspects regarding assumptions on reality and knowledge that can be recognized behind these premises and the practice of SD.

"Real" World and Presentationalism

The traditional distinctions of *ontology*, *epistemology*, and *methodology*, form the habitual framework to drive a philosophical discussion. But the isolation of these issues may not be the most adequate or clear strategy. Given the interrelated nature of such categories and the misleading discussion on those terms which is present in a large part of organization and management science literature – part of this confusion will be exposed and clarified below – then a different plan will be used to develop the rest of this article. An instinctive option is to pick up significant issues and to relate them with what we can identify as part of the core of SD.

Where to start? Assumptions concerning a "real" world can be a first step to take since the traditional debate developed through the years around the claim of building models of "social systems" has fueled part of the discussions in systems science. As most of examinations on philosophical matters the debate has been permeated by a confusion originated in terms and words without the proper examination of the topic. However, this short revision is useful for opening the assessment of the premises of SD regarding a real world.

A prominent example is the criticism made by influential commentators who state that SD models represent an assumed "objective" real world [49]. This kind of critique usually labels SD as a "hard" approach, e. g. [48], meaning with such a term models of an assumed objective machine-like world – and habitually including and inverting the meaning of the term "positivism" – a mistaken assessment that still can be seen nowadays, e. g. [16]. This type of comments can be illustrated with the following quote from the work of Flood and Jackson [23]:

> System dynamics models still center on capturing the structure of the "real world" ... the underlying assumption of SD that there is an external world made up of systems the structure of which can be grasped using models built upon feedback processes ... Because intentions derive from inside social systems, from the conscious human actors which constitute them, many possible appreciations of the nature and purpose of particular social systems are possible. SD simply does not deal with the innate subjectivity of human beings ... In essence the argument is that social systems cannot be studied, in the way of system dynamics, objectively from the outside (pp. 78, 79–80).

However, that is not what SD looks for. The mentioned emphasis on the examination of human decision-making processes and on the assumptions behind, the notion of mental model, the fact that model significance rests on its suitability for a particular purpose, among other aspects, should suffice to illustrate the point. Already Forrester in 1961 [24] emphasized: "a model can be useful if it represents only what we *believe* to be the nature of the system under study ... we are forced to commit ourselves on what we believe is the relative importance of various factors. We shall discover inconsistencies in our basic assumptions ... Thorough any of these we learn" (pp. 57–58, emphasis original). This should be enough to discard the assessment made by Flood and Jackson, and similar critics, who mistakenly placed SD in the terrain of a sort of naive realism as depicted in the quote above; this point is extensively clarified by Lane [55,56]. More importantly, this argument is helpful to introduce the discussion about the nature of SD models, the assumptions behind these models, and their relation with a "reality". The notion that drives these matters has been labeled in SD literature as "mental model" and even though this concept is not free of debate [21] it is placed at the core of the discipline.

The idea of mental model was already addressed by Forrester, as it has been indicated; he demarcated it in 1970 as the mental image of selected concepts and relationships of the world around us that we carry in our heads [26]; furthermore, "the mental model is fuzzy. It is incomplete. It is imprecisely stated ... [It] changes with time" (p. 213). Doyle and Ford [22], looking for a consensus on this subject, propose as a definition: "a relatively enduring and accessible [conscious], but limited, internal conceptual representation of an external system (historical, existing or projected) whose structure is analogous to the perceived structure of that system" (p. 411). Sterman [98] summarizes what the expression "mental model" refers to: "our beliefs about the networks of causes and effects that describe how a system operates, along with the boundary of the model (which variables are included and which are excluded) and the time horizon we consider relevant ... Most of us do not appreciate the ubiquity and invisibility of mental models, instead believing naively that our senses reveal the world as it is. On the contrary, our world is actively constructed (modeled) by our senses and brain" (p. 16–17). It should be noticed that these mental models may refer as well to planned or desired systems existing in the mind of the modeler [54]. The ultimate goal of system dynamics is to enhance our learning processes by testing and improving our mental models in a way that becomes consistent with the complexity of the systems that we face and design everyday [98].

Then, what is a SD model as related to some "real" world? Or how is it related to the notion of mental models? The usual option to answer these questions is to frame the discussion in the debate realism/anti-realism. This examination is even more relevant considering that some criticism labels SD as anchored on "realism". A brief clarification follows.

Realism? Anti-Realism?

Typical of latest debates in history of science concerns the dispute between the so-called *realism*, i. e. theories are true *or* false as descriptions of the world, and *instrumentalism*, i. e. theories are more or less adequate. This latter position is closer to pragmatism: theories are just instruments to systematize – and for many philosophers predict – observations, but theories are not claims about the world; or they can be also subjected to linguistic frameworks but still always truth-value-less. An overview of this debate is made by Leplin [60]. Yet, here it should be clarified that both positions and those definitions – habitually taken by historians of science – are two sides of the same coin. In short, both are forms of idealism. This assessment will be commented next.

On the one hand, scientific "realism" is usually defended using the method of abduction as source of knowledge, i. e. inference to the best explanation, which is just a form of induction and hence it is confirmationist, ultimately relativistic. It is not realism at all; it is exactly the opposite, i. e. idealism. For instance, take the influential ideas of Sellars [91] who defends an illustrative traditional position of scientific realism holding the source of knowledge on observation and relying on justification by induction for building theoretical frameworks; he emphasizes: "Laws in question are stipulated to be inductively established in the observation framework" (p. 313), the theories are then refined by empirical generalizations and what he calls further "injections" of images of the theory into the observational framework, in a typical process of instructional correction. A further refined and formal model of such a framework (proposed by Friedmann, holding a model–submodel relation) is commented by Morrison [66] who still holds, nevertheless, the search for truth and justification as support for his criticism and the necessity of confirmation [67]; see also for example the paper of Kukla [52] with a criticism to the ideas of Friedmann, yet also supported also on confirmation. In short, this "realism" is just idealism as we know it. Similar "realist" positions abound on the literature of the history of science. E.g. Smith [95] postulates a realist stance based on "common sense" but, since for him "science begins with observations" (p. 53), his realism ends up, indeed, trapped in sensations, i. e. idealism. Quantum mechanics does not escape the debate, e. g. arguing in favor of realism within the subjective theory of probability (the Copenhagen interpretation, Bohr, Heisenberg, etc.), Dickson [20] stands on the verification criterion for discussing on what he calls "quantum realism"; naturally such base can not succeed, which is anyway best self-explanatorily reflected in the first part of the title of his paper "An Empirical Reply to Empiricism"; simply there is no such reply. Another case is the criticism of Brown [13] to the deterministic notion of "realism" that Cherniak discards (who supports it on computer simulation of finite agents); the criticism of Brown shows a Galilean notion of realism, i. e. achievable true descriptions of the world via laws, but in a very "complex world" and thus, for him, inaccessible to agents with limited cognitive capacity; this Galilean view is found in large part of complexity science. In other attempts, for instance Schlagel [90] defends "contextualistic realism" which ends up in relativism: elements of the world have a conditional status relative to contexts and conditions, the existents are real relative to the particular structures and contexts on which they depend; his proposal is just idealism and can be better described as a sort of multi-phenomenalism, indeed relativism. The diverse "realist" positions *actually rooted on non-negotiable empiricism* seem endless; e. g. further examples are found in [1,59,75], and in most papers with the expression "scientific realism" in the abstract.

On the other hand, let us consider a supposedly opposite position, for instance instrumentalism. Instrumentalism actually ends up in relativism as well, e. g. with respect to the points of view where instruments can be applied; at the end the notions of "applicability" or "adequacy" become reference frameworks for establishing partial truths. Newton-Smith [68] proposes a conciliatory position, he calls it "modest realism" which ends up indeed in a sort of "moderate" relativism; a catalog of various scientific realisms can be found in that paper too, all of them addressing still the question of how such truths about a real world can be sustained (justified). In short, instrumentalists declare the dependence on observations, i. e. idealism, and thus this position is in fact sharing assumptions with the alleged "realists". Leplin [60] summarizes: "some theory can be reduced to observation by defining or translating theoretical terms into terms that describe observable conditions. The remainder must be construed instrumentally" (p. 394).

The "realism" vs. instrumentalism debate is futile and yet it is perhaps one of the core discussions in philosophy of science. But unlike these influential historians of science, the presented dispute tends to be dismissed by several professional philosophers who see it as self-serving and unsophisticated; Fuller [33] underlines this assessment though he also illustrates the consequences of leaving such pointlessness discussions to endure. In this case, two seemingly unaware idealist factions argue about the best way to establish positive knowledge, e. g. either with a supposedly "true" description of the world (more precisely: *phenomena*, subjectivism) or by pragmatism (and again: *phenomena*, relativism). This shared *idealism* is the matter of the next section.

Presentationalism

A broader debate can be assessed framed in the opposition between *realism* and *idealism*, see e. g. [72] and [77]. It will be shown that SD rests on the broad stance known as idealism, i. e. presentationalism – this latter term is preferred here just for clarity [10].

Firstly, it should be commented the widely inverted and misleading use of both terms. And there is good company. Blackmore [10] shows various celebrities that became misusers of the expressions in question such as the former president of the American History of Science Society and prominent Harvard University professor, Erwin Hiebert, and Sir Russell Brain, outstanding neurologist and former president of the British Association for the Advancement of Science. Add the seemingly customary tendency to quote references in second hand without inspecting direct sources and a few decades later we have the terms used in exactly their opposite original sense in journals and books. Blackmore pictures the situation:

> Like-minded 'empiricists' have restricted what they understand by the term to what idealistic philosophers of science *have wanted them to understand by it*. And since the term 'realism' sounds good to 'tough-minded empiricists' and since idealistic philosophers such as Hume, Comte, Schuppe, and Mach and their recent successors *scarcely if ever have admitted their idealism*, many scientists and historians of science have let themselves be seduced into reversing the normal epistemological definitions of 'realism' and 'idealism'. Even worse, respected scholars such as Stillman Drake, perhaps our most outstanding authority on the manuscripts of Galileo, and Larry Laudan, a young, energetic, and much published commentator on Mach and 'empiricism', have allowed themselves to become advocates of hopelessly naive 'non-philosophical' positions. Drake is sure that Galileo held no philosophical position, or that if he did, it had no effect on his scientific work. Laudan is equally positive on the basis of Mach's written comments that his phenomenalistic epistemology had no influence on Mach's 'empirical' methodology of science. The simplicity of their views can party be explained by their tendency to understand by the term 'philosophy', not one's most basic universal assumptions, *but expressed talk about speculative matters*. Similarly, many 'materialists' who feel sure that the physical world is directly given in experience, and who accept the idealist Kant's distinction between 'science' and 'metaphysics', are convinced that anyone who identifies the physical world with what is *beyond* immediate experience is an 'idealist' and 'metaphysician' and (following Wittgenstein) 'is merely uttering nonsense' (p. 131 in [10]).

In order to clarify, it is appropriate to underscore the attitude behind an empiricist epistemology. A major defining posture is what has been labeled as *idealism*, given the natural disbelief of a world beyond senses, which is the pillar of an empiricist epistemology. The term "idealism" comes from the "idea" of Bishop Berkeley, who took physical objects as "ideas" which included sensations and thoughts: "It is evident to any one who takes a survey of the objects of human knowledge, that they are either ideas actually imprinted on the senses, or else such as are perceived by attending to the passions and operations of the mind, or lastly ideas formed by help of memory and imagination, either compounding, dividing, or barely representing those originally perceived in the aforesaid ways." [8]. This idealism relies entirely on senses and mind-dependent worlds since consequently sense-data were the only things of whose existence our perceptions could assure us, and that to be known is to be 'in' a mind, and therefore to be mental. Berkeley, therefore, concluded that nothing can ever be known except what is in some mind, and that whatever is known without being in my mind must be in some other mind [85]. Hunter [47] summarizes:

> As a result of their constraints on knowledge and meaning, empiricists tend to be skeptical of necessary truths that are independent of mind and language, and of putative eternal abstract entities (p. 110).

This idealism can be equally identified with terms such as 'phenomenalism', 'neutral monism', or 'subjective idealism' (e. g. [10]), or presentationalism. In other words, "anything in time or space, anything than can be known by the human mind, is phenomenal" (p. 146 in [15]).

Taking this posture to the context of science, Bartley [4] provides the implications: "Presentationalists see the subject matter of science not as an external reality independent of sensation. The subject matter of science is our sensory perceptions. The collectivity of these sensations is renamed 'nature' … The aim of science is seen not as the description and explanation of that independent external reality but as the efficient computation of perceptions … [It] became the dominant twentieth-century *philosophy* of physics" (p. 11, 16, emphasis original). In general this position is the pillar of the prevalent conception of science which has been fueled by physics (for instance in the interpretation of quantum mechanics of Bohr and Heisenberg) and backed by influential names like Mach, Russell, Wittgenstein, Ayer, Lewis, Carnap, etc.

To appreciate the contrast, a standard definition of *realism* can be:

> 'Realism' … is used for the view that material objects exist externally to us and independently of our sense experience. Realism is thus opposed to idealism, which holds that no such material objects or external realities exist apart from our knowledge or

> consciousness of them, the whole universe thus being dependent on the mind or in some sense mental. It also clashes with phenomenalism, which, while avoiding much idealist metaphysics, would deny that material objects exist except as groups or sequences of sensa, actual and possible (p. 126 in [10]).

Apart from the particular emphasis on material objects (as opposed to Berkeley's *ideas*), another point is that realism defends a cosmocentric thesis opposed to the anthropocentric view in the discussions of the alleged "realists" of science presented earlier; in the latter case the observer-centered learning process is fundamental, and it is carried on via induction looking for acquiring positive, verifiable, and true knowledge – or justified true belief. Moreover, let us recall that historians of science denote with the term "realism" just the concern with supposed true descriptions of the world. But in fact *realism* does not imply that knowledge is achievable, it does not imply that the world is a perfect clock, it does not imply determinism, it does not imply that there can be correspondence between theories and such real world; these are different affirmations that unfortunately seem to be muddled inside the same bag. One thing is to assume a real world beyond senses. But a different inquiry is the character we ascribe to it. Another different issue is the role we assume for our senses. Another very different concern is the question of knowability, etc. Rescher [77] illustrates typical examples of the confusing use of terms in literature: "The three positions to the effect that real things just exactly are things as *philosophy* or as *science* or as '*commonsense*' takes them to be – positions generally designated as *scholastic*, *scientific* and *naive* realism, respectively – are in fact versions of epistemic idealism exactly because they see reals as inherently knowable and do not contemplate mind-transcendence for the real" (p. 187).

Coming back to presentationalism, this position then assumes that we are imprinted by the environment, and we call to this *impressions* "knowledge". Given the limitation of our senses then presentationalism postulates that nothing more can be known; and thus, such assumption is used to construct the world in our minds: for a presentationalist, strictly speaking, the world is not re-presented since we do not have access to it, the world is just what is *presented* to our senses: the world as we experience it happens to be the world itself; and, since anything that can be known by the human mind is, then, phenomenal (sensations, etc.), therefore knowledge strictly depends on – and is source in – what is sensed, e. g. observed. Hence knowledge needs to be justified in order to avoid error; and yet, the only existent knowledge is the imperfect evidence sourced in sensation and, nevertheless, a foundation that can be justified is pursued. This is the popular plan we have come up with, so far, to try to avoid the destruction of empiricism made by Hume. The picture can be summarized:

> Almost all traditional epistemologies are Lamarckian in their accounts of the growth of knowledge. This is conspicuously true of presentationalism, almost all adherents to which maintain an inductivist, justificationalist account of knowledge growth, according to which knowledge is constructed out of sensations (as building blocks or elements) by a relative passive process of combination, accumulation, repetition, and induction (p. 25 in [4]).

This position is identified with *idealism* and most popularly associated with *epistemological empiricism*, with all its assumptions and its consequent scientific method.

The last point to underline is that this epistemology has subordinated ontology. A remarkable illustration of this type of problems was already made by Bowman [11]: "The result … is the more or less deliberate abnegation of a genuine epistemology and the substitution for it of a highly formal logic. Hence the paradox illustrated equally in the case of Plato and, recently, of Mr Russell, of a radical empiricism (expressed in Plato's Protagorean theory of sensation and in Russell's subjectivism) subsisting side by side with the extremist rationalism. Such a dualistic position is the despair of philosophy, and indicates a failure in the synthetic work of thought" (p. 485). This despair is easy to recognize in current science which presents a contradictory ontological position whose difficulty is found in its subsumption under a radical epistemology. Indeed extreme empiricism, i. e. presentationalism, has become the metaphysics, i. e. the theory of reality, of our science. The debates on "scientific realism" presented earlier picture this failure. On this particular subject the different use of terms in literature is a source of confusion; but here the terms have been inverted by historians of science and scientists; and beyond a semantic confusion this has brought a narrow conception and a very restricted examination of epistemological assumptions. In short, the subjectivism of Descartes and Kant – or more precisely, Kantian idealism – is what now is labeled as "realism", e. g. *everything* has become "phenomena". As a matter of fact common expressions like "objective phenomena" or "real phenomena" uncover an idealistic position where the "objective" or "real" are just *phenomena*. Indeed, the so-called scientific "realism" commented above is nothing more than *a sort of empiricism* driven by Hume and Kant. This "realism" is just ontology overlapped by epistemology (*idealism*). More precisely, regarding Kant, let us recall his "Transcendental Idealism" which was the common answer of Kant when he was accused of idealist, e. g. see [93], denying to be a "dogmatic idealist" (in the Berkeley sense; see e. g. [101]); a full discussion of this failed defence of Kant is made by Guyer [37]. In particular Turbayne [101] defends Kant when he was accused of having misinterpreted (or even completely having misunderstood) Berkeley's idealism; yet, Turbayne's conclusion summarizes the Kantian idealism (and its ambiguity) unmistakably: "The Kantian antidote to this is not the a priori nature of space, but *its reality or subjectivity*, which assimilates space and its contents into the realm of ideas, and thus *prevents* illusion" (p. 243, emphases added). Regarding consequences, perhaps the best summary is the radical idealist position of Mach – who denied even the existence of atoms since they cannot be observed. Kant seems to be taken for granted without a proper reflection on his position.

These few points were commented since this debate is the major informer of the method and the assumptions of management science, organization science, and social science in general, places where SD has its roots. More important, by making this clarifications then a clear ground for SD can be envisaged. It can be affirmed that SD has been mistakenly labeled as "realist" by many commentators alluding the alleged aim of building "true" descriptive theories of the world, and using the misleading definitions of historians of science. But also from the discussion above it should be clear that SD safely rests on presentationalism. Moreover, the identified presentationalist stance known as "instrumentalism", and in general the so-called anti-realists positions (independent of the inverted use of the term), fit to the SD worldview: the abstractions from our experience are arranged in mental models which form knowledge that we want to improve in order to make better decisions. The SD models built for achieving this goal are judged against their adequacy and suitability for a particular purpose; these models are not claims about the world but instruments for systematizing observations and for boosting learning processes using experimentation via simulation.

The Discussions on Positivism: Presentationalism and Knowledge

An issue previously mentioned which is closely connected with presentationalism is relativism and positivism. Since positivism usually – and mistakenly – is pejoratively associated with a supposed objective representation of reality, then an important clarification is needed. In short: positivism is consistent with presentationalism and with relativism as well.

Presentationalism Brings Positivism

Blackmore [10] reminds that strictly speaking neither "rationalism" nor "empiricism" are properly epistemological terms at all; the entrenched idealism has led just to this narrowed identification. For instance, usually the term "empiricism" is synonymous of knowledge sourced in observation, i. e. in a restricted epistemological context, but such popular narrowness is inaccurate. Blackmore remarks: "Granted, that if one *means* by 'empiricism' not just an extensive and careful concern with empirical evidence but *restricting* reference or knowledge or both to sensory appearances, then there are indeed epistemological implications. One has become an epistemological phenomenalist or subjective idealist, or if you will, a positivist" (p. 130, emphases original). This clarification is needed for two reasons; on the one hand, as it was stated, SD has been labeled as "positivist" but the critics take this term as a sort of naive realism; the confusion is patent once we realize that positivism actually is a consequence of idealism, the opposite doctrine of realism. On the other hand, since SD is better identified with idealism then a sort of positivism can be also associated with it, but not the sort of "positivism" that the critics have in mind but the authentic positivism; and yet we will see that with the use of simulation positivism does not necessarily fit either.

The case of management science is a good example regarding the discussion on relativism and positivism. Let us consider the traditional and unfortunate sharp division between "hard vs. soft" which also takes the form "quantitative vs. qualitative". However, this discussion is misleading as well. It is not difficult to find researchers that claim to be anti-positivists but being themselves grounded on positivism (e. g. empirical observation, verification, induction, etc.) without noticing the contradiction. A good example is the claimed opposition between positivism and phenomenology. Yet, phenomenology is authentic positivism when is committed to evidence – Husserl himself underlines this aspect – see e. g. [94]. Indeed it is easy to appreciate an inverted use of the term "positivism" in literature; in management science this is a favored and widespread practice where so-called anti-positivists do not notice their positivism. The fact is that anthropocentrism is the ground in our most influential epistemologies that recognize the obvious imperfections of our sensorial apparatus but, nevertheless, rely knowledge on sensation (observation, etc.), that is, positivism, which is nothing more that our anxiety to confirm and validate, i. e. justify, our "imperfect" knowledge, e. g. empirically. This is a simplification of highly loaded terms; clarifications and further discussions can be found elsewhere, e. g. regarding positivism see [9,97,100].

Positivism is Anchored on Justification

Within a presentationalist worldview the search for confirmation and verification is nothing less than the search for justification of knowledge where the intellectual authority lies in sense experience. From a presentationalist account it is straightforward to have a justificationist approach for confirming and verifying theories. Following Bartley [4]:

> Preoccupied with the avoidance of error, they suppose that, in order to avoid error, they must make no utterances that cannot be justified by – i. e., derived from – the evidence available. Yet sense perception seems to be the only available evidence … The claim that there is an external world *in addition* to the evidence is a claim going *beyond* the evidence. Hence, claims about such realms are unjustifiable. Crucial to the presentationalist argument are, then, two things: the desire to give a firm foundation or justification to the tenets of science, and the construal of sense experience as the incorrigible source of all knowledge (pp. 12–13, emphases original).

In fact justification philosophy taken as the search for epistemic 'authorities' has been the dominant style of western philosophy looking for "well-grounded" knowledge. For instance in the customary view of knowledge as *justified true belief*, e. g. in the sense of Russell [86]) – as the result of systematic analysis "of our sensory experience of a knowable external reality" (p. 47 in [96]). Within this popular position the central problem of epistemology – as succinctly formulated by Radnitzky [76] – becomes:

> "When is it rational or, so to speak consistent with one's intellectual integrity, to *accept* a particular position?" The formulation suggests the direction in which the answer is to be sought: "When concerned with a statement, a theory, etc., accept those and only those statements, theories, etc., which not only are true but whose truth has been established" (p. 282, emphasis original).

The goal of justification is usually entrenched within the method of induction where every new repeated observation is a confirmation that validates – justifies – the theoretical statement. Even with weaker conditions the way of reasoning is the same, for instance within the ideas of Ayer where strict verifiability is seen as a too rigid criterion – he introduces *confirmability to some degree*, instead of complete and conclusive verifiability (see [88]); justification is still pursued. The appeal of justification can be explained because it looks for avoiding relativism (inherently attached to presentationalism) since not all positions are equally good or bad and it suggests to look for something beyond blind belief [76]. Though it is not the only option, nevertheless, it is the most common view of science; the concerns on validation, justification, verification, confirmation, and generalization, are part of this popular and influential view. Here the observer is the fixed point of reference. In short, within a justificationist logic, it is rational to accept only those positions that have been justified according to the rational authority which in the case of presentationalism is sense experience, consistent then with the highly influential ideas of Locke, Berkeley, Hume, Mach, Carnap that have shaped our prevalent view of science [4].

Justification in System Dynamics

Turning back to SD, it must be recalled the role of mental models whose characteristic nature of "abstractions based on experience" can be better assessed within a presentationalist stance. Here justification has also a place. How is this knowledge justified? Already Forrester [24] emphasized, within the debate of model validation, that, "knowledge of all forms can be brought to bear on forming an opinion of whether or not a model is suitable to its particular purpose" (p. 129). Therefore, Forrester [25] also emphasized that "we can never prove that any model is an exact representation of 'reality' … Models are then to be judged, not on an absolute scale that condemns them for failure to be perfect, but on a relative scale that approves them if they succeed in clarifying our knowledge" (p. 3–4). This sort of relativism will be addressed next.

With the aim of placing this stance within the discussion of history of science, Barlas and Carpenter [3] addressed the "philosophical roots of model validation" associating SD with what they called a "relativist philosophy of science". In this view justification is pursued: such a knowledge is seen as socially, culturally and historically dependent and it becomes socially justified belief. Here "a valid model is assumed to be only one of many possible ways of describing a real situation … for every model carries in it the modeler's world view … validation is a matter of social conversation" (p. 157). Hence, confirmation and verification are pursued through a social process relative to a frame of reference. This sort of moderate relativism was later criticized by Vásquez, Liz and Aracil [103] for whom such relativism is unacceptable in spite of their recognition that there is no privileged single model (or set of models); since they are also concerned with epistemological justification these authors present Putnam's internal "realism" as a more adequate way to conceptualize the type of knowledge consistent with the assumptions about reality held by SD practitioners; in short, these authors underline that mental models are the source of knowledge – and its justification – helping to select the structures that must be assumed as working in real systems; here knowledge is taken as internal to the conceptual scheme of SD. Since there can be many models for a given situation, the authors argue that this framework gives the possibility of convergence as a result of "the strong interactive character of mental models" (p. 34) recognizing that in any case SD modeling is a process of revision and adjustment. With this proposal these authors seek to achieve justification and some realistic representational content in spite of the plurality of alternative SD models available for a specific situation. It is easy to see that this proposal is still relativistic, in this case knowledge is relative to the mental models and the conceptual scheme – though the mentioned authors would not agree since for them there is a "reality" given by the internal representational schema, in this case the mental models of the modelers.

The fail to recognize presentationalism as the epistemology which is driving ontology is at the root of the presented discussions. This is a distinctive trait present in large part of the philosophy of science literature. However, the characteristic problem of mistaking positivism with a sort of supposed "objectivism" is also present in this discussion – in fact, Barlas and Carpenter, following the misdirecting literature on the subject, argue that the relativist philosophy that they defend rejects positivism; Vásquez, Liz and Aracil also follow the same inertia. However, these theses, which seek for confirmation, verification, justification, reflect the search for *positive* knowledge within an idealist epistemology. In the first case, knowledge is confirmed and accepted through social interaction and it is relative to a context. In the second case, knowledge is justified on mental models. To appreciate a genuine contrasting position see an anti-justificationist approach to validation in computer simulation in [51]; the core of anti-justificationism can be found in the work of Popper, e. g. [73,74].

System Dynamics as Presentationalist

It should be clear that the assumptions on SD reject naive realism; the models are not supposed to be accurate and corresponding descriptions of an external true reality – and furthermore, this is not what the term "positivism" refers to. In any case validation in SD does not mean a supposed "positive proof" or to assess the development of "true models" of the world. On the contrary, SD aims at enhancing our ways of reasoning, it emphasizes a process of learning so as to consistently improve our mental models which are product of our experience and the operations of our mind. Hence SD is closer to presentationalism. Knowledge can be socially justified and our mental models can be enhanced. Moreover, a particular emphasis on the modeling *process* is also underlined – see e. g. [29,44,89] and it will be commented in the eight section devoted to simulation.

A Brief Note Regarding Social Theory

Another important discussion is related to the theories about the "social world" that are supposedly held in SD, i. e. the social theory behind SD if any. Part of the misguided debate is explained by the widespread use of the traditional framework of Burrell and Morgan [14] whose oversimplification of social science in four paradigms has deviated major issues covering important topics under a too practical and inadequate schema – e. g. see a criticism in Deetz [19]. Jackson [49] provides a summary of such usual misconstruction:

> System dynamics … is essentially functionalist in nature. It sees system structure as the determining force behind system behavior … If humans are free to construct social systems as they wish, what determining influence does system structure have? … This tension between determinism and free will is unresolved (p. 39).

It should suffice to recall from the discussion above that the *aggregate* approach of SD is not a theory of human behavior; SD is not concerned with *individual* action. Furthermore, it does not assume that a structure, of any kind, determines human behavior either, i. e. the sort of determinism that Burrell and Morgan [14] oppose to "free will" and in the lines of the already vague term known as "structuralism" – see an early clarification of the problems of such type of oversimplification in [84]. In any case, this sort of criticism has been answered and clarified by Lane [56] who has underlined the main point: SD is concerned with aggregate social phenomena and not with individual meaningful actions [55]. Moreover, system dynamics does not propose invariant causal laws, as Lane [56] also concludes: "The only universal law/theory on offer is a grand methodological, or structural theory, associated with a representation scheme … it does not attract the determinism-related criticisms attached to grand theory in the sense of Parsons and Mills" (p. 111). In [57] Lane proposes to link system dynamics with a different framework: agency/structure theories.

Servomechanism

Part of the confusion is because of the misunderstanding by various commentators of the notion of feedback that underlies SD. This point has been also a source of misconception given the use of feedback in theories of control applied to social systems, e. g. cybernetics. Consider for example the following comment by Flood and Jackson [23]:

> The attempt of SD thinkers to model external reality is misguided … The emphasis placed on "structure" as the means of revealing knowledge about the optimal behavior of systems cannot be accepted … SD modelers using feedforward control appear to believe that there are optimal future states that we should steer systems towards" (pp. 80, 81).

Such criticism apparently is associated with the notion of feedback used in cybernetics. However SD does not pursue optimization, let alone by studying "knowledge revealing" structures in order to achieve supposed optimal behavior patterns. Instead, the goal is to have a better understanding of feedback structures in order to enhance decision-making and policy design. This point has been also addressed by Lane [53,55]. The central clarification of this issue has been made by Richardson [78] who distinguishes two different threads in the development of the concept of feedback in the social sciences: the cybernetics thread and the servomechanisms thread. The failure in noticing these two different lines of thought has produced various misconceptions regarding the notion of feedback in SD, a concept that has been shown as one of its building blocks. On the one hand, the cybernetic conception of feedback is defined in terms of input and output, it is limited only to loops of negative polarity which in turn are conceived as the mechanisms of control – and hence there is a particular interest in goal seeking and goal formulation given the concern in cybernetics for achieving adaptive behavior via directed processes and homeostatic mechanisms; feedback mechanisms guide this pursue of viable behavior which is carried by goal-seeking processes. On the other hand, in SD, coming from the servomechanisms thread, feedback loops are taken as intrinsic parts of the system (and not just as mechanisms of control), it includes also loops of positive polarity, and such feedback structures are seen as responsible for counterintuitive behaviors and policy resistance in social systems; here the analysis is directed toward policy design.

Explanation and Mechanism

The next interesting question would be how we can achieve better understanding and better policy design by enhancing our mental models. How can we characterize this type of knowledge?

A solid account of *explanation* should be placed at the heart of any scientific activity. The general inquiry about [scientific] explanation has to do with "learning how the process of doing science facilitates understanding, and what type(s) of understanding science provides" ([7] p. 307). In a very intuitive way, a first approach to explanation might be associated plainly with removing puzzlement [6]. It is also common to affirm that an explanation aims to answer queries of *why* in order to provide understanding [87]. Yet, to characterize such idea is a major and open unresolved question in philosophy of science [69]. The notion of causality has traditionally played a central role; this view has pervaded most of scientific research where theory development and explanations are essentially conceived as the search for causes, e. g. [50,87]. However, several explanations are not essentially based on simple causal relations but on other approaches such as identification, models, analogies, formal linguistics, laws of association, laws of co-existence, variational principles, among others [83]. Berger [7] underlines the prominence of this question when attempting to characterize the explanations provided by nonlinear dynamical modeling:

> Mathematical modeling [is recognized] as one of the central activities of science, and it is reasonable to say that modeling explanations dramatically increase our understanding of the world. But the modeling explanations found in contemporary scientific research show that the interesting claims of causal accounts are untenable … An adequate account of scientific explanation must accommodate modeling explanations, because they are simply too central to ignore (pp. 329–330).

The main goal of this section is to explore the position and the characterization of the kind of explanation pursued in system dynamics. This characterization fits with the presented core of SD and the presentationalist stance. And again various clarifications will be needed along the way.

Causality

The difficult issue of causality can be treated in several senses. In the first place, a possible relationship associated with the term "determinism" on human behavior is dismissed with a previous argument: SD causal models does not point at supposed laws of causality governing human action [55,56]. What is more interesting is to investigate the concept of causality as such in SD models; after all, a large part of SD modeling relies on what are known as "causal"-loop diagrams. Forrester emphasized the term *interrelationships* [24] where feedback loops are understood as closed informational paths connecting in a sequence decisions that control actions [25]; he labeled it as a "circular cause-and-effect structure" (p. 1–9). In fact the development of *causal*-loop diagrams has become important in SD practice; in particular, flow diagrams were initially recognized as useful pictorial representations that help to formulate and communicate the structure of a dynamic model [24]. These models are ultimately theories of behavior, surely causal theories of behavior. But consistent with presentationalism, these theories are sourced in the mental models of the modeler, there is no direct connection to an alleged causation in a real world. Sterman summarizes a definition: "a causal diagram consists of variables connected by arrows denoting the causal influences among the variables" [98]; here, every link represents *what the modeler believes is* a causal relationship between two variables; this causal attribution is seen as a central feature of mental models, as Sterman also stresses "we all create and update cognitive maps of causal connections among entities and actors … Within a causal field, people use various cues to causality including temporal and spatial proximity of cause and effect, temporal precedence of causes, covariation, and similarity of cause and effect" [98]. Again, the core of the discussion should be driven by the concept of mental model in order to deliver a clear discussion on this view of causality held in SD. This section outlines a framework.

Most of the time we seem to hold a strong causal view of the world. In particular, the causal relationship – whatever that could be – tends to be the source of explanatory power, i. e. the *explanans*, and one usual source of validity, that is, *for having a relevant valid explanation we must have a causal relationship*. That is the usual principle, and it is usually associated to the term "determinism". Recalling the discussion on presentationalism, one should note that – as Hesslow [42] underlines – if we are going to retain a Humean view of the world, i. e. consistent with idealism, then it seems that we have two different paths, probabilistic or deterministic. The latter one is of interest here (for the probability account of causation see e. g. [43]). Within a deterministic approach, a cause is always sufficient condition or a part of a sufficient condition for the effect, that is, if A_t is a nonsufficient cause of $B_{t'}$, then there must be some auxiliary condition $C_{t''}$, [with $t'' < t'$], such that A_t in combination with $C_{t''}$ is sufficient for $B_{t'}$. Hesslow clarifies this "sufficiency principle" [42]:

> The deterministic approach has been something of a received view of causation. This view, which we may call the 'sufficiency principle' is also common among scientists. The sufficiency principle is not in itself strictly deterministic. It does not mean that every event has a sufficient cause, only that if an event has a cause, then it has a sufficient one. However, it seems that the popularity of the sufficiency principle is a reflection of a widely spread, though usually implicit, commitment to the stronger thesis, that every event has a sufficient cause (p. 592).

The cited paper of Hesslow aims to show that the deterministic approach is superior to the probabilistic one, that is, the idea that the probabilistic account presupposes determinism. Indeed what is happening is the trap of Hume – so to speak. The issue in hands is illustrated with the *Humean fork*: based on observation of constant conjunction of events – altogether with temporal priority, i. e. the cause is observed prior to its effect – we *suppose* a causal connection between them – see summaries in e. g. [46,70] and the original work of Hume [45]. These *suppositions* are arranged in our mental models, that is, in our theories about what we assume as the relevant causal connections that we suppose so as to explain the world.

With this framework in mind, the network of causal connections that the SD modeler believes to be relevant indicates a sort of "sufficiency principle", but of a special kind. It is sufficient relative to the purpose of the model, as it has been indicated above. And it is sourced in the mental model since SD is seen as a vehicle for learning and not as a device for operating a "real world". How are these causal relationships portrayed? In a generic form, a feedback loop based stance is based on the fact that decisions in a time t affect the environment which affects again the future assessment of the situation in a time $t'(t' > t)$ which usually is the base for new upcoming decisions and actions taken in a time t'' $(t'' > t')$ and so on. These relationships are not necessarily close in space or time. Furthermore, a related issue is what can be labeled as "multiple causality". The complexity of social systems is a current concern for social scientists; the multiple interactions among several agents, actors, or entities, is what has been distinguished as the key to study and to understand complex systems because of the recognition of our inability to deal with them based on traditional incomplete simple-causality thinking – the assumption that explanation of phenomena can be satisfactory or even sufficient based in simple unidirectional causal relationships between variables or constructs. What is more, feedback loops have important implications associated with counterintuitive behavior that usually we do not consider easily or that we misunderstand; they have a key role in complex settings; in fact they are, to a large extent, responsible for the arising of complex behaviors; this is a central affirmation in SD [26,98]. The simplest example might be the tendency that we have to infer linear growth from single first order positive feedback loops. But the behavior here actually is exponential. Let S be the state of a system and g the constant fractional growth rate, the linear first order positive loop and its solution are:

$$\frac{\mathrm{d}S}{\mathrm{d}T} = gS$$

and it has as solution:

$$S = S_0 \mathrm{e}^{gt} .$$

The central question is: are these causal theories, portrayed in system dynamics, claims about the world? i. e. Are these models assumed true descriptions of the world? Clearly no. From the presentationalist stance of SD, the causal-loop diagram is essentially what the modeler *believes* is the relevant causal network for the problem in hands. It constitutes his theory about it. The source of knowledge is the mental model and causality is only a supposition of the modeler and a way to arrange knowledge, it is not an affirmation of truth about a supposed causal world. And furthermore, causation as such is not the source of explanation provided by SD. In order to clarify this we should take a look to the notion of explanation held in system dynamics.

Mechanism

System dynamics aims to answer *why* questions. This is done generally via the development of *dynamic hypotheses*. A core premise of SD has always been to enhance learning and to provide understanding [30]. This aim has been stated from the very beginning; for instance Forrester asserted in *Industrial Dynamics*: "Our objective is to enhance understanding and to clarify our thinking about the system" p. 57 in [24]).

How is this goal pursued? A SD model should be able to account for a specific problematic behavior which is explained in terms of the structure of the model; here the term "structure" refers to the stock and flow organization, the feedback loops and the rules of interaction [98]. This approach to explanation is known as a dynamic hypothesis and is the core concept in order to provide understanding from a system dynamics point of view; its endogenous character is the chief feature that makes it intelligible; for instance: "One key task in this search for insightful, system level understanding is the telling of 'system stories' – coherent, dynamically correct explanations of how influential pieces of system structure give rise to important patterns of system behavior" (p. 1 in [65]). In fact this task represents one of the more significant research lines [79].

How can we characterize this particular kind of explanation? As a first point, the notion of organized social complexity helps to drive this discussion. A quote borrowed from Hayek illustrates it:

> Where we have to deal with such social wholes we cannot, as we do in the natural sciences, start from the observation of a number of instances which we recognize spontaneously by their common sense attributes as instances of 'societies' or 'economies' ... What we group together as instances of the same collective or whole are different complexes of individual events, in themselves perhaps quite dissimilar, but believed by us to be related to each other in a similar manner: they are classifications or selections of certain elements of a complex picture on the basis of a theory about their coherence (p. 43 in [40]).

Based on the quote above of Hayek, he suggests conceiving the *explanation* as modeling [104], and for social sciences he rejects the usual prediction and control aspirations and asks the reader to focus more on models to explain typical processes [64], he depicts it with biology: "It deals with pattern-building forces, the knowledge of which is useful for creating conditions favorable to the production of certain kinds of results, while it will only in comparatively few cases be possible to control all the relevant circumstances" (as cited in p. 202 in [104]). Hayek calls it "explanation of the principle". Essentially he means the explanation of a *kind* of phenomena instead of particular events. As another example consider mathematics: "A set of equations which shows merely the form of a system of relationships but does not give the values of the constants contained in it, is perhaps the best general illustration of an explanation merely of the principle on which any phenomenon is produced" (p. 291 in [39]). This analogy illustrates the notion of abstract relations that would build an "explanation of the principle" which can be associated with *mechanism* as the source of explanatory power – instead of causality.

Again, a clarification is needed given the widespread identification of the term "mechanism" with ontic commitments. Fundamentally mechanism is a kind of *explanation*. It should be noticed that "mechanism" refers to *epistemological* issues. However, the term is habitually associated with assumptions about reality. But Hogben already clarified in 1930:

> In any discussion between the two [mechanist and holist or vitalist], the combatants are generally at

> cross purposes. The mechanist is primarily concerned with an epistemological issue. His critic has always an ontological axe to grind. The mechanist is concerned with how to proceed to a construction which will represent as much about the universe as human beings with their limited range of receptor organs can agree to accept. The vitalist or holist has an incorrigible urge to get behind the limitations of our receptor organs and discover what the universe is really like (1930, p. 100, as cited in [12], p. 347).

The explanatory notion of mechanism is well underlined by Grene: "Let us look for a mechanism which might underlie the phenomena we hope to understand, seeking wherever we may relevant sources from which to derive ... an analogue of a possible mechanism ... [Such an explanation is of value because it tells us] *how in fact those phenomena are produced*" (as cited in [12], p. 346, emphasis original).

However, there is still no agreement about what a mechanism is and how it appears to succeed in science as a way to provide understanding. Perhaps the most complete account is the one of Tabery [99] who proposes integrating two complementary points of view. These two aspects are (i) the interactions among several parts and, (ii) the activities associated with these interactions. Both characteristic are taken as necessary for having a mechanism-based explanation.

On the one hand there is the emphasis on interactions, a thesis supported by Glennan [36] with a central concern on the nature of *complex systems* since the role of parts and its interactions are conceived as indispensable. The work of Bechtel and Richardson [5] develops the association of mechanism and complex systems (in biology and psychology) and emphasizes the tasks of decomposition and localization as the heuristics in order to uncover mechanisms. This position replies to the conventional view of a mechanism as merely the interactions between causal processes as the essential *explanans*; Glennan [36] stresses: "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (p. 344). Glennan clarifies also that he avoids using the term "causal-law" and instead uses "change-relating generalizations" because these relations are not exception-less as traditionally a law is understood. In addition, he emphasizes the very different character of such account which is in opposition to the traditional causal view in which mechanisms are sequences or chains of events leading up to a particular event – which is often associated in systems theory literature with "linear thinking".

On the other hand we have activities. Machamer, Darden and Craver [62] emphasize this aspect adding that mechanisms are not only inter-connected entities but also activities producing regular changes from initial to finish conditions; they call themselves dualists since for them both notions – entities and activities – are necessary to constitute a mechanism:"The organization of these entities and activities determines the ways in which they produce the phenomenon. Entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration" (p. 3). It is important to underline their critique to Glennan's view arguing that the concept of *activity* is fundamental to understand the changes produced (because of the activities) through the process and not only as the black-box view of change of states or change of properties of the inter-connected entities; they picture it clearly with the following statement: "it is not the penicillin that causes the pneumonia to disappear, but what the penicillin does" (p. 6). Furthermore, in order to account for a mechanism they emphasize three distinctions: set-up conditions (as *part* of the mechanism, not as a sort of input; this includes relevant entities and their properties and initial states), intermediate activities (including also relevant entities, properties, and an intelligible account of the activities that link them) and termination conditions (such as privileged endpoints, equilibrium states or the final stage of some unitary integral process). They also draw attention to the fact that mechanisms take place in nested multi-level hierarchies and that they usually are not full pictures but truncated abstract accounts – *a mechanism schema* – depending on the required level of detail or aggregation.

System dynamics modeling is perhaps one of the best ways to picture this kind of explanation. One can distinguish two main components in the structure of SD models: the physical and institutional assumptions – including the chosen parts/variables and the interconnections between them, and the decision rules of the agents [98]. The interplay between the physical structure and the associated decision rules as the explanation for behavior is a foundational aspect. Indeed the interconnections and the activities needed to account for a mechanism are included in these system dynamic models structures. Specifically, the activities producing change can be referenced in the links of the models and the decision rules describe how the interactions produce certain activities. The whole set of initial and final conditions and the inter-connected parts engaged in producing activities characterize a mechanism.

For example, the simplest mechanism is perhaps a single feedback loop. The set-up conditions are the initial values of the variables involved, the termination condition is the endpoint of the loop which can be accounted in a mechanism as "the final stage of what is identified as a unitary, integral process"(p. 12 in [62]). The intermediate activities are depicted by the links and the application of the decision rules. For instance, a simple positive first order loop can produce exponential behavior beyond the particular values of the variables involved. More intricate structures are the source of different behaviors, i. e. the change in the values or patterns of variables through time.

This stance fits an explanation of an abstract principle in the sense of Hayek. The structure is the source of explanation of patterns of behavior, i. e. the change in the values or patterns of variables of interest through time. This is known in SD as a dynamic hypothesis [98]. With this focus on aggregate patterns – instead of individual events – as consequence of the structure, it can be said that the "causal" mechanism is indeed the loop structure of the system, or the particular and relevant feedback substructures of the model that may explain the behavior; for instance Richardson [78] illustrates it in this way: "The 'cause' of an arms race is viewed not as a given event or even a given sequence of events, but as a feedback structure dominated by self-reinforcing positive loops, within which events take place" (p. 338). These types of explanations are based on *mechanisms* as explanatory power and not in simple causal relationships as the source of explanation, even less in (substantivalist) causality, i. e. change in singular properties/entities. Furthermore, these hypotheses are developed for each problem consistent with the mental models of the modelers. This is why system dynamics is not committed to specific theories and only to the explanation of problematic behaviors in terms of structure of the model in order to enhance learning and decision-making.

A particular remark must be made. It can be noticed a natural link between mechanism and the idea of "generic structure". This latter expression has been used in different senses in SD literature. Lane and Smart [58] trace the evolution of this concept – see also [71], they identify three different interpretations. One of these view cannot be connected with mechanism, the one popularized by Senge, e. g. [92,109,110], usually known under the expression "system archetypes". The lack of computer simulation within this interpretation points at a problem of validity in its scope and claims, as Lane and Smart discuss [58], e. g. since this perspective skips the possibility of formal computer model building then the relation between structure and behavior is weak and the mentioned approach becomes just a hasty shortcut from problematic behavior to insights and principles, and without the experimental spirit of SD for enhancing learning. But two other notions are relevant. On the one hand, generic structures can be conceived as general models (theories of behavior) of a class of systems that are associated with a domain of application, e. g. urban development, supply chain, economic growth. Lane and Smart label these structures as "canonical situation models". On the other hand, a generic structure may refer to theories of mathematical structures (feedback loops, levels, rate equations, etc.) that generate corresponding dynamic behaviors, i. e. "systems belong to the same class if they can be represented by the same structure … This dynamic structure when abstracted from any application domain data defines the class of system" (p. 93 in [58]); therefore they offer transferability of structure across diverse domains. These models can be labeled as "abstracted micro-structures", e. g. patterns of exponential growth, goal seeking, oscillation, etc. [58,98]. Both interpretations look for establishing a general class of models that formally link structure with behavior and they constitute an important line of research. These developments contribute to different aspects of SD practice; in particular they directly fuel the processes of conceptualization and formal model construction (e. g. see [31,58,98]), and more important, they enhance understanding and the improvement of our mental models as long as we can exploit the powerful idea of having general classes of models, either within a domain of application or across different domains by transferring structures across them.

Consequently, system dynamics explanations can be characterized as *mechanisms*, since there can be found the source of explanatory power. In spite of its causal diagrams, the *explanans*, i. e. that which does the explaining, is based on mechanisms – dynamic hypotheses based on structures; and the problematic behavior is the *explanandum*, i. e. that which is explained. Following Glymour: "Remains, however, a considerable bit of science that sounds very much like explaining, and which perhaps has causal implications, but which does not seem to derive its point, its force, or its interest from the fact that it has something to do with causal relations (or their absence)" (as cited in [p. 212 in 83]). The theories built with SD are essentially structure-based and not content-based (substantivalist) explanations i. e. they are not associated with intrinsic properties of objects or entities but with the consequences of processes and activities entrenched in relevant parts of the structure of the model. Recalling Hayek who identifies explanation with modeling, it is interesting to notice the range of his thought expressed half a century ago and that accurately illustrates SD explanations:

> Any model defines a certain range of phenomena which can be produced by the type of situation which it represents. We may not be able directly to confirm that the causal mechanism determining the phenomenon in question is the same as that of the model. But we know that, if the mechanism is the same, the observed structures must be capable of showing some kinds of action and unable to show others (p. 221 in [38]).

A further reminder follows. Since SD was previously identified with instrumentalism, then a mechanism is not to be taken as a description; here a mechanism is an instrument for arranging observations. It should be kept in mind that mechanism is a kind of explanation which refers to epistemological issues.

The popularity of simple causality as the way to characterize the explanation of phenomena contrasts with the assumptions made in SD: structures that generate processes responsible for behavior. This is consistent with the purpose of system dynamics simulation which might be oriented to activities such as theoretical-representations building, articulation and testing in order to learn in and about complex systems [98]. System dynamics uses simulation as a method which is different from the traditional inductive logic of research that deals with single instances which attempts to confirm theories via repeated observation. However, SD does not dismiss presentationalism as it was shown. This should be highlighted as an important and distinctive characteristic of SD. Though there is a commitment with a real world, justification is rooted in social processes and on mental models, and it is also relative to the purpose of the model. Furthermore, the goal is to enhance our decision-making processes by improving our mental models. How can we characterize such method? A short comment follows.

Simulation and Method

In a plain sense "simulation means driving a model of a system with suitable inputs and observing the corresponding outputs" (p. 23 in [2]). But simulation actually is not just a matter of number crunching. Its scope is broader. And it represents another challenge for philosophy of science. Winsberg [107] illustrates it:

> Typically, to a philosopher of science, epistemological issues arise when we try to justify high level theoretical claims based on low level data or specific observational reports. But simulation is about starting with theory and working your way down. This kind of epistemology is, to the philosopher of science, a curious beast. It is an epistemology that is concerned with justifying inferences from a theory to its application – an inference that most philosophy of science has assumed is deductive and consequently not in need of justification (p. S447).

What is simulation in SD? It can be affirmed that it is a technique able to represent and test theoretical concepts and not only – in a narrow sense – a tool to just solve mathematical problems. Besides, the emphasis on processes, on patterns of collective action and on the relations between components and its dynamic consequences can be better addressed with simulation because of its capacities to represent these issues with fewer restrictions than other approaches [35]. But there is more. Simulation reflects a very different attitude. This way of inquiry suggests a whole different and new scientific methodology [82,106,107]. Winsberg emphasizes that "simulation represents an entirely new mode of scientific activity – one that lies between theory and experiment … a form of theory articulation or 'model building' (pp. 117, 119 in [106]). Axelrod [2] indeed suggests "a third way" of doing science:

> Simulation as a way of doing science can be contrasted with the two standard methods of induction and deduction … Simulation is a third way of doing science. Like deduction, it starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively. Unlike typical induction, however, the simulated data comes from a rigorously specified set of rules rather than direct measurement of the real world (p. 24).

Moreover, its strength rests on the capacity for conducting *experiments* [82]. This emphasis on experimentation is the key to understand why this approach is different. Our mental models are nothing more than theoretical models that attribute properties and relations to the systems they represent; the relevance of these theoretical models depends on the purpose of the model. And computer simulation simply permits experiments of these (theoretical) models. This is where the novelty and the power of this methodology are to be found, in the very iterative process of model building and experimentation via simulation. This position contrasts with the traditional prominence of assumed representational capacities of *theories* and *models* where usually the emphasis has been placed, see [107]. But computer simulation has a distinct epistemology [105] that emphasizes *the process of modeling*. The method was demarcated by Forrester [24] in *Industrial Dynamics*:

> Simulation consists of tracing through, step by step, the actual flows of orders, goods, and information, and observing the series of new decisions that take place … This is the counterpart of trying a new policy or organizational structure in the real system… After a simulation run comes interpretation of the results. Did it turn out as expected? If not, why? As the experiment is examined, new questions arise … This is a process of invention and trial … Each simulation result teaches, and it also prompts additional questions … Such experimentation will yield new insights into the characteristics of the system that the model represents (pp. 23, 44–45, 55).

Hence, the method of simulation through continued experimentation is aimed at providing better understanding of the modeled system. As it was mentioned, the method calls attention to the *process* – see also [44,89]. Indeed SD aims at developing a *modeling* culture (consistent with [80]) that gives emphasis to model building as an ongoing dialectic between stakeholders instead of a mapping exercise concerned on the efficacy of the model itself.

Why is fundamental the use of the computer? Perhaps the best answer is provided again by Forrester [26]:

> We stress the importance of being explicit about assumptions and interrelating them in a computer model … The most important difference between the properly conceived computer model and the mental model is in the ability to determine the dynamic consequences when the assumptions within the model interact with one another. The human mind is not adapted to sensing correctly the consequences of a mental model … The computer model … is a statement of system structure. It contains the assumptions being made about the system … Generally, the consequences are unexpected (pp. 213–215).

The shortcomings of our mental models coupled with the complexity of the systems we model lead to the use of the computer. Sterman [101] summarizes these drawback with aspects such our flawed cognitive maps and our erroneous inferences about dynamics. As it was shown, the strength of explanation and understanding in SD is not in the causal models as such; the heart lies in the development of dynamic hypotheses with the use of simulation in order to enhance our understanding and our decision-making processes. Explanations are posed under the notion of mechanism; but this is an iterative process that seeks a central aim: to improve our own theories about the world.

Future Directions

There are several aspects to develop based on this reflection on various assumptions behind system dynamics. The central argument was built around the position known as presentationalism. This stance integrates and informs many of the debates on related subjects, e. g. validation, and it characterizes the initial purposes and assumptions held in the field. However, there are a number of lines to emphasize and to develop.

We have focused on the role of the idea of "mental model" for the practice of SD. Yet, it can be seen utilization of SD models under assumptions which are closer to a naive realism that seem to ignore the purposes of enhancing understanding and learning processes. This article should help to underscore that system dynamics is less naive – and hence more powerful – when we recognize a presentationalist stance which means that our theories about the world are just that, *theories* based on our experience and on the operations of our own mind. We can improve these theories with the use of system dynamics, that is, making our assumptions about systems explicit and using simulation as a method for enhancing understanding and for developing explanations that guide our processes of systems design. This recognition includes the relative nature of justification of knowledge held in SD and the emphasis on mechanism as a powerful way to develop explanations about complex systems.

It has been introduced mechanism as the way to characterize the particular type of explanation pursued in SD. The explanatory force does not rest on causal relations as such but on the structures – physical and decision rules aspects – and on the dynamic processes and activities that explain change. The idea of mechanism is shaped in SD under the expression "generic structure". The focus on understanding behavior in terms of abstract structures is a central line of inquiry. This article underlines the importance of developing such a line of research since it has been located at the core of the kind of knowledge that SD provides. The issue of unification to provide understanding of diverse phenomena is a definitive step in the way to assert that the field progresses as long as broader range of phenomena may be explained with the same mechanism. Should be the advance of system dynamics assessed by the progress in this type of study? Behind this discussion there are major and provocative issues that arise to be developed.

The long debate of qualitative and quantitative modeling is informed by this characterization. It is clear that the issue of explanation compels theorists and practitioners to ask themselves what is the kind of explanation they are pursuing and if such explanations are enough and satisfying; it is worthy to ask for qualitative modeling what kind of understanding it gives and how it can be characterized, in other words, to give an account of *explanans* and *explanandum*. Another line to develop might be oriented around the following question: what would be *essential* criteria for comparing different arranges or modes of organization in order to identifying them as belonging to the same type of mechanism? The identification and search of mechanisms becomes a powerful heuristic for guiding the modeling process.

There are promising suggestions for philosophy of science as well. SD offers guidelines, e. g. can the ways in which system dynamicists work provide meaningful insights, or even concrete accounts, for the philosophical unresolved issue of explanation? The mechanism depicted in system dynamics proposes a kind of explanation that goes beyond the received view based on causation. A related question is whether explanation must always follow a deductive path; the classic models of Hempel, e. g. [41], emphasized the condition of deduction and general laws for having an explanation; however, the explanation in SD is not framed under a deductive schema from universal covering laws, instead it can be conceived as a sort of abductive reasoning based on the understanding of the dynamics of the model as a way to understand the actual behavior it accounts for. A further issue is that in spite of the lack of universality, i. e. no universal laws, SD models aim to provide understanding for a diverse range of phenomena that might share relevant influential structures and similar associated behaviors, that is, it accounts for regularities in order to unify them in a certain kind of explanation under the same explanation. This situation insinuates a flavor of paradox because of the traditional rigid association of universal laws with the explanation of regularities, but in SD there are no general laws though the aim is to explain general regularities. The study of *models* and computer simulation – instead of abstract theories and traditional methodologies – is an additional indication that SD suggests for philosophy of science, including the emphasis on the modeling process.

Bibliography

1. Aronson JL (1988) Testing for Convergent Realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 188–193
2. Axelrod R (1997) Advancing the art of simulation in the social science. In: Conte R, Hegselmann R, Terna P (eds) Simulating Social Phenomena. Springer, Berlin, pp 21–40
3. Barlas Y, Carpenter S (1990) Philosophical roots of model validation: two paradigms. Syst Dyn Rev 6:148–166
4. Bartley III WW (1987) Philosophy of biology versus philosophy of physics. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 7–45
5. Bechtel W, Richardson RC (1993) Discovering complexity: Decomposition and localization as strategies in scientific research. Princeton University Press, Princeton
6. Benjamin AC (1941) Modes of scientific explanation. Philos Sci 8:486–492
7. Berger R (1998) Understanding science: Why causes are not enough. Philos Sci 65:306–332
8. Berkeley G (1948–1957) The works of George Berkeley, Bishop of Cloyne. Thomas Nelson and Sons, London
9. Black M (1934) The principle of verifiability. Analysis 2:1–6
10. Blackmore J (1979) On the inverted use of the terms 'Realism' and 'Idealism' among scientists and historians of science. Br J Philos Sci 30:125–134
11. Bowman AA (1916) Kant's phenomenalism in its relation to subsequent metaphysics. Mind, New Ser 25:461–489
12. Brandon RN (1984) Grene on mechanism and reductionism: more than just a side issue. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol II: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 345–353
13. Brown HI (1990) Cherniak on scientific realism. Br J Philos Sci 41:415–427
14. Burrell G, Morgan G (1979) Sociological paradigms and organizational analysis. Heinemann, London
15. Chapin JP (1941) Idealism and its relation to science. Philosophy of Science 8:142–146
16. Checkland P, Pidd M, Morecroft J (2004) Working ideas, insights for systems modeling: The Broader community of systems thinkers. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the 22nd International Conference of the System Dynamics Society, Keble College, University of Oxford, England. System Dynamics Society, Albany
17. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: some research questions. Syst Dyn Rev 16:225–244
18. Coyle RG (1979) Management system dynamics. Wiley, Chichester
19. Deetz S (1996) Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy. Organ Sci 7:191–207
20. Dickson M (1995) An empirical reply to empiricism: Protective measurements opens the door for quantum realism. Philos Sci 62:122–140
21. Doyle JK, Ford DN (1998) Mental models concepts for system dynamics research. Syst Dyn Rev 14:3–29
22. Doyle JK, Ford DN (1999) Mental models concepts revisited: some clarifications and a reply to Lane. Syst Dyn Rev 15:411–415
23. Flood R, Jackson M (1991) Creative problem solving. Wiley, Chichester
24. Forrester JW (1961) Industrial Dynamics. Press MIT, Cambridge
25. Forrester JW (1971) Principles of Systems. Wright-Allen Press, Cambridge
26. Forrester JW (1975) Counterintuitive behavior of social systems. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 211–244

27. Forrester JW (1975) Industrial Dynamics: A Major breakthrough for decision makers. In: Collected papers of Jay W Forrester. Wright-Allen Press, Cambridge, pp 1–29
28. Forrester JW (1975) Industrial Dynamics – After the first decade. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 133–150
29. Forrester JW (1985) "The" model versus a modeling "process". Syst Dyn Rev 1:133–134
30. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3:136–149
31. Forrester JW (1994) System dynamics, systems thinking, and soft OR. Syst Dyn Rev 10:245–256
32. Forrester JW (2003) Dynamic models of economic systems and industrial organizations. Syst Dyn Rev 19:331–345
33. Fuller S (1994) Retrieving the point of the realism-instrumentalism debate: Mach vs. Planck on science education policy. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 200–208
34. Gershenson C, Aerts D, Edmonds B (2007) Worldviews, science and us. Philosophy and complexity. World Scientific, Singapore
35. Gilbert N, Troitzsch KG (1999) Simulation for the social scientist. Open University Press, Buckingham
36. Glennan SS (2002) Rethinking mechanistic explanation. Philos Sci 69:342–353
37. Guyer P (1983) Kant's intentions in the refutation of idealism. Philos Rev 92:329–383
38. Hayek FA (1955) Degrees of explanation. Br J Philos Sci 6:209–225
39. Hayek FA (1942) Scientism and the study of society, Part I. Economica, New Ser 9:267–291
40. Hayek FA (1943) Scientism and the study of society, Part II. Economica, New Ser 10:34–63
41. Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175
42. Hesslow G (1981) Causality and determinism. Philos Sci 48:591–605
43. Hitchcock CR, Salmon WC (2000) Statistical explanation. In: Newton-Smith WH (ed) A Companion to the philosophy of science. Blackwell Publishers, Malden, pp 470–479
44. Homer JB (1996) Why we iterate: scientific modeling in theory and practice. Syst Dyn Rev 12:1–19
45. Hume D (1740) A treatise of human nature. Oxford University Press, Oxford
46. Humphreys P (2000) Causation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 31–40
47. Hunter B (1992) Empiricism. In: Dancy J, Sosa E (eds) A Companion to Epistemology. Blackwell Publishers, Oxford, pp 110–115
48. Jackson M (1991) Systems methodology for the management sciences. Plenum Press, New York
49. Jackson M (2003) Systems thinking: Creative holism for managers. Wiley, Chichester
50. Jobe EK (1985) Explanation, causality, and counterfactuals. Philos Sci 52:357–389
51. Kleindorfer GB, Ganeshan R (1993) The philosophy of science and validation in simulation. In: Evans GW, Mollaghasemi M, Russell EC, Biles WE (eds) Proceedings of the 1993 Winter Simulation Conference. IEEE, Piscataway, pp 50–57
52. Kukla A (1995) Scientific realism and theoretical unification. Analysis 55:230–238
53. Lane D (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10:101–134
54. Lane D (1999) Friendly amendment: A commentary on Doyle and Ford's proposed re-definition of 'mental model'. Syst Dyn Rev 15:185–194
55. Lane D (2000) Should system dynamics be described as a 'Hard' or 'Deterministic' systems approach? Syst Res Behav Sci 17:3–22
56. Lane D (2001) Rerum cognoscere causas: Part I – How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate? Syst Dyn Rev 17:97–118
57. Lane D (2001) Rerum cognoscere causas: Part II – Opportunities generated by the agency/structure debate and suggestions for clarifying the social theoretic position of system dynamics. Syst Dyn Rev 17:293–309
58. Lane D, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12:87–120
59. Leplin J (1992) Realism and Methodological Change. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 435–445
60. Leplin J (2000) Realism and Instrumentalism. In: Newton-Smith WH (ed) A Companion to the Philosophy of Science. Blackwell Publishers, Malden, pp 393–401
61. Luna-Reyes LF, Andersen DL (2003) Collecting and analyzing qualitative data for system dynamics: methods and models. Syst Dyn Rev 19:271–296
62. Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philos Sci 67:1–25
63. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the system dynamics method. Productivity Press, Cambridge, pp 23–57
64. Milford K (1994) In pursuit of rationality. A note on Hayek's The Counter-Revolution of Science. In: Birner J, van Zijp R (eds) Hayek, Co-ordination and Evolution. Routledge, London, pp 323–340
65. Mojtahedzadeh M, Andersen D, Richardson GP (2004) Using digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20:1–20
66. Morrison M (1988) Reduction and realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 286–293
67. Morrison M (1990) Unification, realism and inference. Br J Philos Sci 41:305–332
68. Newton-Smith WH (1988) Modest realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 179–189
69. Newton-Smith WH (2000) Explanation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden/Oxford, pp 127–133
70. Newton-Smith WH (2000) Hume. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 165–168
71. Paich M (1985) Generic structures. Syst Dyn Rev 1:126–132

72. Pettit P (1992) Realism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 420–424
73. Popper K (1963) Conjectures and refutations. The growth of scientific knowledge. Routledge and Kegan Paul, London
74. Popper K (1968) The Logic of Scientific Discovery. Hutchinson, London
75. Psillos S (1996) Scientific realism and the "pessimistic induction". Proceedings of the Biennial Meeting of the Philosophy of Science Association, Part I: Contributed Papers. University of Chicago Press, Chicago, pp S306-S314
76. Radnitzky G (1987) In defense of self-applicable critical rationalism. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 279–312
77. Rescher N (1992) Idealism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 187–191
78. Richardson GP (1991) Feedback thought in social science and systems theory. Pegasus Communications, Waltham
79. Richardson GP (1996) Problems for the future of system dynamics. Syst Dyn Rev 12:141–157
80. Richardson KA (2002) On the limits of bottom-up computer simulation: Towards a nonlinear modeling culture. In: Sprague RH Jr (ed) Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03). IEEE
81. Roberts EB (1978) System dynamics – An introduction. In: Roberts EB (ed) Managerial applications of system dynamics. Pegasus Communications Inc., Waltham
82. Rohrlich F (1990) Computer simulation in the physical sciences. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 507–518
83. Ruben D (1990) Explaining explanation. Routledge, London
84. Runciman WG (1969) What is structuralism? Br J Sociol 20:253–265
85. Russell B (1912) The problems of philosophy. Oxford University Press, Oxford
86. Russell B (1948) Human knowledge: Its scope and limits. Simon and Schuster, New York
87. Salmon W (1992) Explanation. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 129–132
88. Salmon WC (2000) Logical Empiricism. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 233–242
89. Schaffernicht M (2006) Detecting and monitoring change in models. Syst Dyn Rev 22:73–88
90. Schlagel RH (1981) Contextualistic realism. Philos Phenomenol Res 41:437–451
91. Sellars W (1976) Is scientific realism tenable? PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 307–334
92. Senge P (1990) The fifth discipline: The art and practice of the learning organization. Doubleday, New York
93. Sidgwick H, Caird E (1880) Kant's refutation of idealism. Mind, New Ser 5:111–115
94. Sinha D (1963) Phenomenology and positivism. Philos Phenomenol Res 23:562–577
95. Smith DW (1982) The realism in perception. Noûs 16:42–55
96. Spender J-C (1996) Making knowledge the basis of a dynamic theory of the firm. Strateg Manag J 17:45–62
97. Stace WT (1944) Positivism. Mind, New Ser 53:215–237
98. Sterman J (2000) Business dynamics. Systems thinking and modeling for a complex world. McGraw-Hill, Boston
99. Tabery JG (2004) Synthesizing activities and interactions in the concept of a mechanism. Philos Sci 71:1–15
100. Taube M (1937) Positivism, science, and history. J Philos 34:205–210
101. Turbayne CM (1955) Kant's refutation of dogmatic idealism. Philos Q 5:225–244
102. Vanderminden P (2006) System dynamics – A field of study, a methodology or both. In: Größler A, Rouwette E, Langer R, Rowe J, Yanni J (eds) 24th International Conference of The System Dynamics Society. Radboud University Nijmegen, System Dynamics Society, Albany
103. Vásquez M, Liz M, Aracil J (1996) Knowledge and reality: some conceptual issues in system dynamics modeling. Syst Dyn Rev 12:21–37
104. Weimer W (1999) Hayek's approach to the problems of complex phenomena: an introduction to the theoretical psychology of The Sensory Order. In: Boettke P (ed) The Legacy of Friedrich von Hayek, vol II. Edward Elgar Publishing Limited, Cheltenham, pp 200–244
105. Winsberg E (1999) Sanctioning models: The epistemology of simulation. Sci Context 12:275–293
106. Winsberg E (2003) Simulated experiments: Methodology for a virtual world. Philos Sci 70:105–125
107. Winsberg E (2001) Simulations, models, and theories: Complex physical systems and their representations. Philos Sci 68:S442–S454
108. Wolstenholme EF (1990) System enquiry. Wiley, Chichester
109. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19:7–26
110. Wolstenholme EF (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20:341–356

System Dynamics, Analytical Methods for Structural Dominance, Analysis in

CHRISTIAN ERIK KAMPMANN[1], ROGELIO OLIVA[2]
[1] Department of Innovation and Organizational Economics, Copenhagen Business School, Copenhagen, Denmark
[2] Mays Business School, Texas A&M University, College Station, USA

Article Outline

Glossary

Behavior mode The traditional meaning of the term is the qualitative nature of the observed system behavior, such as damped or expanding oscillations, overshoot and collapse, exponential growth or adjustment to equilibrium, or limit cycles. In linear systems theory, the term has a more specific meaning, cf. the explanation for eigenvalues.

Bode plot (phase and gain plot) A tool used in classical control theory to characterize the frequency response, i. e., the amplification A and phase shift ϕ in the system output variable of interest $x(t) = A \sin(\omega t + \phi)$ compared to the input variable $u(t) = \sin(\omega t)$, as a function of the frequency ω of the input.

Chaos A type of behavior exhibited by nonlinear systems that appears to be approximately periodic but with a seemingly random element. A hallmark of chaotic behavior is that it is sensitive to initial conditions.

Dominant structure A general term for the feedback loops (or possibly external driving forces) that are "most important" in generating a behavior pattern of interest. In nonlinear models, particularly single-transient models, there is frequently a shift in structural dominance, i. e. in the strength and significance of certain feedback loops.

Dynamic decomposition weights (DDW) An application of Eigenvector Elasticity Analysis (EVA) that focuses on how parameter changes influence the relative weights (DDW's) of the system behavior modes in a particular variable.

Eigenvalue An eigenvalue for a square matrix $\boldsymbol{A}$ is a value λ for which the equation $\boldsymbol{A}\boldsymbol{r} = \lambda\boldsymbol{r}$ has a non-zero solution $\boldsymbol{r} \neq \boldsymbol{0}$. The column vector $\boldsymbol{r}$ is called the (right) eigenvector corresponding to the eigenvalue λ. The eigenvalues and eigenvectors determine the behavior modes (components) in the solution to the linear dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$. A real eigenvalue λ leads to an exponential behavior mode $\exp(\lambda t)$ while a complex eigenvalue $\lambda = \tau \pm i\omega$ leads to oscillatory behavior modes $\exp(\tau\, t)\sin(\omega t + \phi)$. The eigenvectors determine the weight, or the degree to which a particular behavior mode is expressed in a particular system variable.

Eigenvalue elasticity analysis (EEA) A method of analyzing the significance of a structural element, say a loop or a link in the model with a gain g, in terms of its marginal effect upon the eigenvalues λ of the system. There are several such measures, such as the *influence measure* $\partial\lambda/\partial g \cdot g$, the *elasticity* $\partial\lambda/\partial g \cdot (g/\lambda)$, or, in the case of complex-valued eigenvalues, the effect upon the damping ratio, natural frequency, damping time, etc., as illustrated in Fig. 7. See also *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Eigenvector See explanation for *Eigenvalue*.

Eigenvector elasticity analysis (EVA) A complement to Eigenvalue Elasticity Analysis (EEA) that looks explicitly at the expression or relative weight of each behavior mode in each system variable. These weights are related to the eigenvectors of the system matrix.

Frequency domain A term used to describe the analysis of signals with respect to frequency. While a time domain graph shows the behavior of the signal over time, the frequency domain graphs shows how much of the signalvariance lies within each given frequency band.

Independent loop set (ILS) Although the number of feedback loops in a model can be very large (theoretically astronomically large), there is a much smaller *independent loop set* that can be considered independent structural elements. For a strongly connected system (where any pair of variables are connected via causal chain in both directions) with N links and n variables, there are exactly $N - n + 1$ independent loops. Simple algorithms exist for constructing independent loop sets, in particular *Shortest Independent Loop Sets (SILS)*. See also explanation for *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Linear dynamical system A system where the rates $\dot{\boldsymbol{x}} = (\mathrm{d}x_1/\mathrm{d}t, \ldots, \mathrm{d}x_n/\mathrm{d}t)$ are a linear function of the state variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ and exogenous or control variables $\boldsymbol{u} = \left(u_1, \ldots, u_p\right)$, expressed by the equation $\dot{\boldsymbol{x}} = \boldsymbol{Ax} + \boldsymbol{Bu}$ where $\boldsymbol{A}$ is an $n \times n$ matrix and $\boldsymbol{B}$ is an $n \times p$ matrix. Unlike nonlinear systems of the general form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, linear systems have analytical solutions based on the eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ (cf. explanation for *Eigenvalues*).

Linear systems theory The mathematical theory of *linear dynamical systems*.

Loop eigenvalue elasticity analysis (LEEA) A form of eigenvalue elasticity analysis (EEA) that uses graph theory to express structural changes in terms of change in the strength of individual feedback loops. *Independent* loops can be assigned individual (loop) eigenvalue elasticities or influence measures just like other structural elements (see explanation for *Eigenvalue Elasticity Analysis (EEA)* and *Independent Loop Set (ILS)*).

Model simplification approach A way of attributing dynamic behavior to particular pieces of structure by replacing the full model with a simplified structure. See also *Structure contribution approach*.

Nonlinear systems Systems of the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ where $\boldsymbol{f}$ is a nonlinear function. See explanation for *Linear dynamical systems*.

Pathway participation metric A measure that decomposes the curvature ($\ddot{x} = \mathrm{d}^2x/\mathrm{d}t^2$) of a variable x_i into the individual driving components, $\ddot{x}_i = \sum_j \partial\dot{x}_i/\partial x_j \cdot \dot{x}_j$. By considering the sign of the curvature relative to the slope, i. e., $\ddot{x}/\dot{x}$, one may define behavior as (apparently) dominated by positive $\ddot{x}/\dot{x} > 0$ or negative $\ddot{x}/\dot{x} < 0$ feedback loops. The component (pathway) with the largest absolute value and the same sign as $\ddot{x}/\dot{x}$ is then defined as the dominant structure.

Quasilinear models Models that are almost linear in structure around the operating point of interest so that they may be well approximated by a linear model.

Quasiperiodic behavior A behavior that is a sum of oscillations of incommensurate frequencies so that the system never returns to exactly the same point (which would be the case for periodic behavior).

Shortest independent loop set (SILS) An *Independent Loop Set (ILS)* that consists of the shortest possible loops (in terms of the number of nodes and links in each loop). Since the choice of ILS is far from unique, an SILS provides a more focused choice of loops, which are typically also easier to interpret due to their short length.

Single-transient models Models where the behavior of interest is the transition toward an equilibrium or constant growth rate. Models are typically nonlinear, exhibit patterns such as smooth transition, overshoot and collapse, growth, or stagnation.

Structure contribution approach A way of linking model structure to dynamic behavior by considering how individual pieces of structure (feedback loops or subsystems) contribute to the behavior pattern of interest by turning the structure on or off (in traditional simulation experiments) or by considering the marginal effect of small changes in structure (the eigenvalue approach). See also *Model simplification approach*.

Definition of the Subject

The link between system structure and dynamic behavior is one of the defining elements in the system dynamics paradigm, yet it is only recently that systematic, mathematically rigorous methods for exploring this link have started to become available. In a sense, a simulation model can be viewed as an explicit and consistent theory of the behavior it exhibits. Although this point of view has certain merits, not least the fact that it lifts the discussion from outcomes to causes of these outcomes and from events to underlying structure [11,59], we are concerned here with a more compact explanation of the system's behavior. In fact, most system dynamics modeling projects report their results in terms of simpler explanations of the observed results, typically in terms of dominant feedback loops that produce the salient features of the behavior.

Most often, dominant structure is thought of in terms of feedback loops and, occasionally, external driving forces to the system. For simple systems with relatively few variables it is usually easy to use intuition and trial and error simulation experiments to explain the dynamic behavior as resulting from particular feedback loops. In larger systems, this method becomes increasingly difficult and the risk of incorrect explanations rises accordingly. There is a need, therefore, for analytical methods that provide some consistency and rigor to this process.

These analytical tools are important to the practitioner because the structure-behavior link is the key to finding leverage points for policy initiatives. And they are important to the theorist because a system dynamics theory of a particular phenomenon is an account of how certain feedback loops cause certain dynamic patterns of behavior to appear. The qualitative understanding of the model behavior is often at least as important as the particular numerical predictions obtained, even in applied studies. Yet the rigor of such an account depends directly on the rigor with which structure-behavior link can be made in a given model.

The classical disciplines of linear systems theory and control engineering have provided a set of concepts and tools, particularly system eigenvalues and eigenvectors, that can also be applied under many circumstances to the nonlinear models found in system dynamics, not as a complete theory but as a pragmatic aid. This article reviews the recent advances in analytical tools based on linear systems theory and discusses its future potential for the both the system dynamics practitioner and the theorist.

Though we strongly believe in the utility of these methods, it is important to realize that advances in nonlinear dynamics and complexity theory in recent decades have shown that it is not possible to construct a complete theory of dominant structure because nonlinear systems are capable of exceedingly complex and intricate behavior that is impossible to predict without actually simulating the system. Furthermore, applications of graph theory to system dynamics models have revealed that the concept of feedback loops has some inherent problems and limitations because there are potentially many different loop descriptions of the same system (see [28,40]). Thus, the analytical tools should be viewed as pragmatic aids to model analysis that can guide the modeler's intuition, rather than universal methods that provide automatic answers.

We first provide a brief historical introduction to the different ways scholars have thought about the notion of dominant structure, including an example of the traditional approach to structural analysis. In the next section we present the formal mathematical representation of linear and nonlinear systems and how one may describe the dynamic behavior in terms of behavior modes and system eigenvalues. In the four following sections we present alternative approaches to performing this analysis. We conclude with a summary of the current state of research and a discussion of future directions.

Introduction

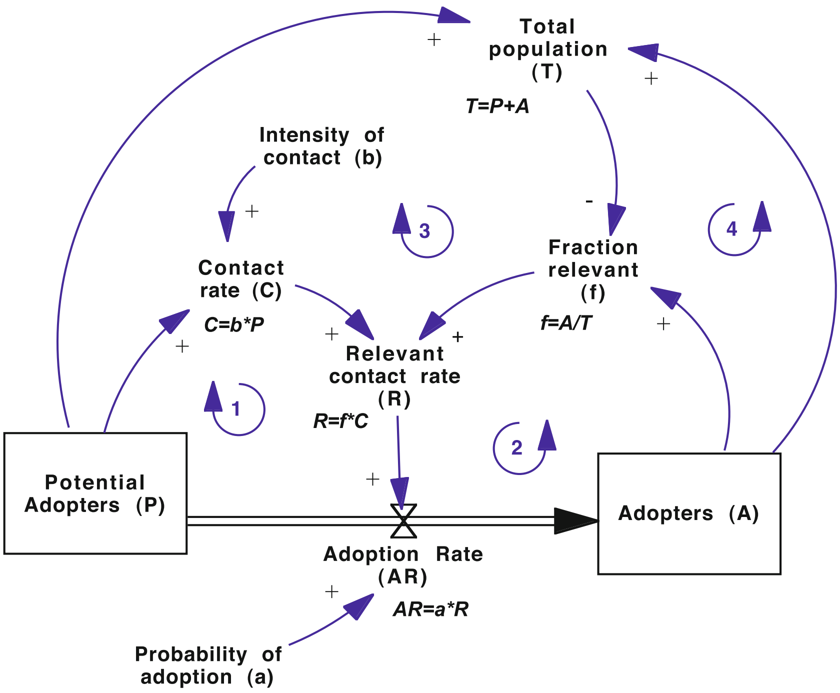
Understanding model behavior is closely related to the process of model testing and validation, for which there is a well-established tradition and an extensive literature in the field (e. g., [2,10,17,36,46,47]). Indeed there is no sharp line between model building, testing, validation, and analysis – in practice, the analyst undertakes all these processes simultaneously [17].

Of particular concern is whether one can identify pieces of structure that are in some sense "important" in generating the observed behavior of interest. Traditionally, system dynamics analysts have relied on trial-and-error simulation to discover these structures, by changing parameter values or switching individual links and feedback loops on and off. The tradition is well developed and includes a set of principles for partial model formulation and testing based the organizational theory of bounded rationality [27,36].

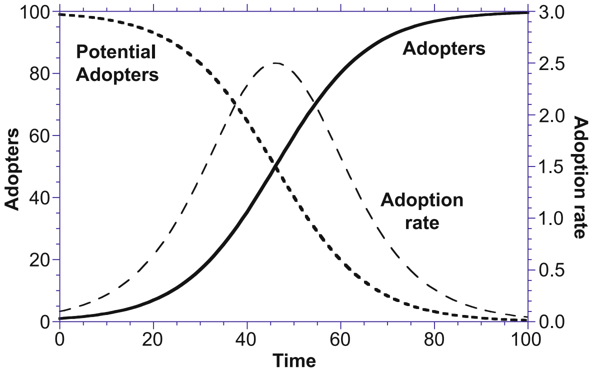
The intuition guiding this effort often relies on simple feedback systems with one or a few state variables, where the behavior is fully documented and understood. In particular, the modeler uses well-understood "generic structures" that seem to appear again and again in system dynamics models, such as "overshoot and carrying capacity collapse", "drifting goal structure", etc. (see [30,56,60] for an account of these structures). Clearly such structures can be a useful aid to understanding if the model is sufficiently simple to allow such simple structures to be identified.

A simple example of a generic structure is the classical model of diffusion, sometimes known as the Bass model ([3], see also Chapter 9 in [59]). The model structure is illustrated in Fig. 1, and the resulting behavior, an s-shaped growth curve, is illustrated in Fig. 2. The idea behind the model is that the adoption of a new technology is driven by the number of users that have already adopted it, through a word-of-mouth effect. One may interpret the s-shaped behavior as the interaction of two feedback loops, namely loop no. 2, the positive "word-of-mouth", and loop no. 1, the negative "exhaustion" loop (see Fig. 1). In the beginning, the positive loop dominates, leading to exponential growth in the number of adopters. Later, however, the negative loop gains strength, and the behavior shifts to an exponential adjustment toward the eventual market saturation. Thus, the traditional feedback loop analysis helps give an intuitive understanding of the dynamics of the model.

In large-scale models with perhaps hundreds of state variables, however, the traditional approach shows significant limitations. In practice, model building and analysis is often done using a "nested" partial model testing approach where one goes from the level of small pieces of structure to entire subsystems of the model, with frequent re-use of known formulations and partial models. Although this approach does carry a long way, it can be very difficult to discover feedback mechanisms that transcend model substructures in ways not anticipated by the modeler in the original dynamic hypothesis. Thus, there is a danger that observed behavior is falsely attributed to certain feedback mechanisms when in fact another set of feedbacks is driving the outcome. Likewise, one may make false inferences about how a particular feedback mechanism modifies the



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 1
The Bass model of diffusion



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 2
Behavior of the Bass model

behavior, e. g., whether it attenuates or amplifies a particular oscillation.

Modern software packages can run extensive tests for sensitivity and "reality checks" where a large number of parameters are varied simultaneously [44]. This is clearly a significant improvement over "manual" trial and error methods, particularly when these methods are combined with statistical inference methods such as Kalman Filtering or Monte Carlo maximum likelihood estimation [6,8,39,43,45,55]. A variant of this approach involves using statistical experimental design and correlation methods to screen for significant model structure (parameters), as suggested by Ford and Flynn [9]. Indeed, the prospects of marrying such methods with modern search and optimization methods like classifier systems [26] or genetic algorithms [19] seem very promising. However, these methods are more addressing issues in estimation, validation and testing than inferences about or understanding how (dominant) structure is causing behavior.

Richardson [47] suggested a taxonomy of approaches to the notion of dominant structure, where he distinguishes along three dimensions, namely linear vs. nonlinear systems, model reduction vs. loop contribution, and the characterization of behavior in terms of time graphs vs. eigenvalues or frequency response. Of these, the distinction between model reduction and loop contribution is the most important.

In the model reduction approach, the idea is to replace a large complicated model with a simplified smaller model that captures the "essence" of the dynamics. A good example of this is Sterman's simple model of the economic long wave [58], which was distilled from the much larger System Dynamics National Model [18]. Eberlein [5,7] attempted to tackle model simplification in a systematic way in linear systems by focusing on retaining specific behavior modes. In large part his results were negative: it is generally not possible to build simpler models that reproduce the salient behavior without sacrificing either the accuracy of the behavior or the ability to relate the simplified model variables to those in the full model. It is fair to say that this line of inquiry has largely been abandoned as a result. Extracting the "essence" of a model remains an art more than a science.

The focus here will be on Richardson's second category, the loop contribution or, more generally, the *structure contribution* approach. It reflects the intuitive idea that if one removes the element under consideration, e. g. by weakening a link or switching off a feedback loop, and the behavior then "disappears", one would say that the element in some sense "causes" the observed behavior.

This notion underlies the traditional trial-and-error simulation approach, sometimes supplemented with methods from the classical control engineering, which focuses on how structural elements modify the behavior of the system, viewed in terms of the frequency response. Typically, the method works "backwards" by starting with simple feedback systems of single loops and then considering the marginal effect of adding links and loops. We discuss this approach in Sect. "Traditional Control Theory Approaches" below.

If, instead, one considers marginal (infinitesimal) changes in structure, e. g. in the strength of a particular link, it is possible to derive rigorous analytical results for the resulting change in behavior expressed as the eigenvalues of the linearized model. One would then say that if a change in a system element has a relatively large effect upon the behavior pattern of interest, this element is "significant" in "causing" the behavior. This is what underlies the *eigenvalue elasticity* and *eigenvector* approaches discussed in Sects. "Eigenvalue Elasticity Analysis", "Eigenvectors and Dynamic Decomposition Weights (DDW)". The marginal and experimental approaches may supplement each other well, where a marginal analysis may identify elements that can then be tested experimentally for their significance.

Unlike the traditional control method and the eigenvalue method that work in the structural and *frequency domain*, the *pathway participation* method (PPM) relates directly to the time path of particular system variables and is more concerned with the qualitative nature of the time path, expressed in terms of signs of the slope (whether growing or declining) and curvature (whether convex or concave) than with numerical measures of degree of influence. Briefly stated, the PPM traces the causal links in the variables influencing the system variable in question and then identifies the most important chain of links. We discuss this method in Sect. "Pathway Participation Metrics".

Common to the approaches discussed here is that they all build upon a precise mathematical characterization of the system behavior. In the next section, we demonstrate how the concepts from linear systems theory may be used to give a precise characterization of behavior in terms of component *behavior modes*.

Characterizing Linear and Nonlinear Systems

A system dynamics model can be represented mathematically as a set of ordinary differential equations

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} \equiv \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) , \tag{1}$$

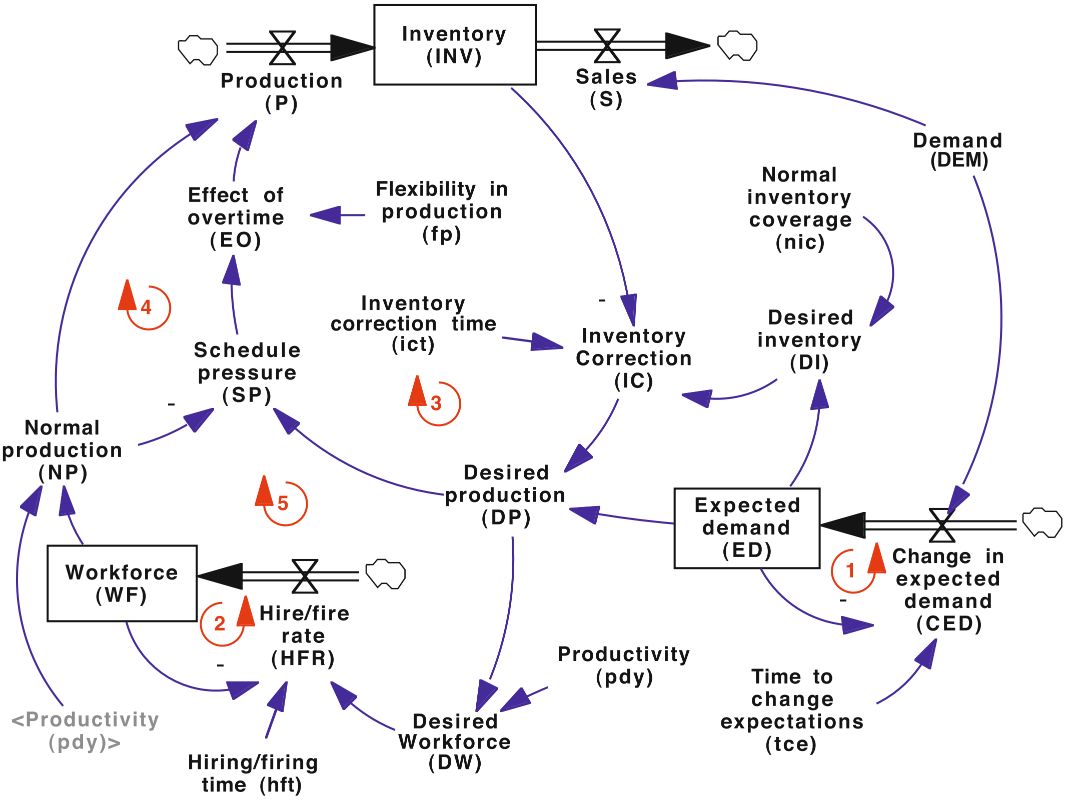
where $\boldsymbol{x}(t)$ is a (column) vector of n state variables (levels) $(x_1(t), \dots, x_n(t))$, $\boldsymbol{u}(t)$ is a column vector of p exogenous variables or control variables $\left(u_1(t), \dots, u_p(t)\right)$, $\boldsymbol{f}()$ is a corresponding vector function, and t is simulated time. In this paper, we restrict our attention to the state variables (levels) of the model for notational convenience, ignoring the auxiliary variables. Mathematically, a model can always be brought to the *reduced* form (1), but in practice, the auxiliary variables give a more intuitive account of the analysis. Likewise, we do not consider time-varying systems (where time t enters as an explicit argument in the function $\boldsymbol{f}$), since these can usually be accommodated by an appropriate definition of the exogenous variables $\boldsymbol{u}$. In general, $\boldsymbol{f}$ is a nonlinear function of its arguments, and we speak of a *nonlinear* system. Conversely, if $\boldsymbol{f}$ is a linear function, we speak of a *linear* system.

Figure 3 and Table 1 show a well-known example, the inventory–workforce model. It has three state variables, Inventory (INV), Workforce (WF), and Expected Demand (ED), and one exogenous variable, Demand (DEM), i. e.,

$$\boldsymbol{x}(t) = \begin{pmatrix} \mathrm{INV} \\ \mathrm{WF} \\ \mathrm{ED} \end{pmatrix} ; \quad \boldsymbol{u}(t) = (\mathrm{DEM}) , \tag{2}$$

and the function $\boldsymbol{f}$ is determined by the equations in Table 1.

Given the model structure (1), knowledge of the initial conditions $\boldsymbol{x}(0)$, and the path of the input variables $\boldsymbol{u}(t)$, the behavior of the model is completely determined. It is in this sense that the model structure (1) constitutes a "theory" of the time behavior $\boldsymbol{x}(t)$, as mentioned in the introduction. Yet, we are interested in methods that yield



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 3
Flow diagram of the inventory workforce model

a more compact explanation, short of having to simulate the entire model structure.

It turns out that in its ultimate form, this dream is beyond reach: Since the days of Henri Poincaré, mathematicians have known that it is impossible to find general analytical solutions to nonlinear systems. Furthermore, the development of nonlinear dynamics and chaos theory has proven that such systems, even when they have very few state variables, can produce highly complex and intricate behavior that goes beyond general analytic methods (e. g., [42,48]). Thus, we will never find a final general theory where we can infer the behavior of the system directly from its structure; instead, we will always have to rely on simulation to discover the dynamics implied by the structure. (This is not to say that no general analytical results exist in nonlinear systems. The field of chaos theory has uncovered a number of universal features, e. g., relating to the transition from periodic or quasi-periodic behavior to chaos, where the transitions show both qualitative and quantitative similarities that are independent of the specific forms of the model equations (see, e. g., [42]). However, these universal features relate to specific situations such as period-doubling or intermittency routes to chaos).

The best we can hope for, therefore, is a set of tools that will guide intuition and help identify *dominant structure* in the model. By dominant structure we mean particular feedback loops that are in some sense "important" in shaping the behavior of interest. To the extent that we can identify such dominant structures, we may say that we have found a "theory" of the observed behavior.

Although the term "behavior" may appear rather loose, experience and reflection tells us that there is a limited number, perhaps a dozen or so, of relevant behavior patterns that dynamical systems can exhibit. Some of these behaviors, like exponential growth, exponential adjustment, and damped or expanding oscillations, are typical of linear systems. Others, like limit cycles, quasiperiodic motion, mode-locking, and chaos, can only be exhibited by nonlinear systems.

Common to the approaches considered in this paper is that they are based on tools from linear systems theory, i. e., they approximate the nonlinear model (1) with a lin-

System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Table 1
Equations of the inventory workforce model

Equation	Name	Units
d/dt(ED) = (DEM − ED)/tce, ED_0 = DEM * df	Expected demand	[Units/Month]
d/dt(INV) = P − S, INV_0 = DI	Inventory	[Units]
d/dt(WF) = HFR, WF_0 = DWF	Workforce	[Workers]
S = DEM	Shipments	[Units/Month]
P = NP * EO	Production	[Units/Month]
EO = fp * (1 − (1 − 1/fp)^SP)	Effect of overtime	[Dimensionless]
NP = WF * pdy	Normal production	[Units/Month]
DI = ED * nic	Desired inventory	[Units]
SP = DP/NP	Schedule pressure	[Dimensionless]
DEM = 1(Exogenous)	Demand	[Units/Month]
DP = ED + IC	Desired production	[Units/Month]
IC = (DI − INV)/tci	Inventory correction	[Units/Month]
HFR = (DWF − WF)/hft	Hire/fire rate	[Workers/Month]
DWF = ED/pdy	Desired workforce	[Workers]
hft = 5	Hire/fire time	[Month]
pdy = 1	Productivity	[Units/Month/Worker]
tce = 4	Time to change expectations	[Month]
fp = 1.05	Flexibility in production	[Dimensionless]
nic = 3	Normal inventory coverage	[Months]
ict = 2	Inventory correction time	[Months]
df = 0.5	Disequilibrium fraction	[Dimensionless]

earized version, using first-order Taylor expansion around some operating point $\boldsymbol{x}_0, \boldsymbol{u}_0$, i. e.,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u} - \boldsymbol{u}_0) , \quad (3)$$

or, by redefinition of the variables $\boldsymbol{x} \to \boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) \times (t - t_0)$ and $\boldsymbol{u} \to \boldsymbol{u} - \boldsymbol{u}_0$,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) , \quad (4)$$

where $\boldsymbol{A}$ is constant $n \times n$ matrix of partial derivatives $\partial f_i/\partial x_j$ and $\boldsymbol{B}$ is constant $n \times p$ matrix of partial derivatives $\partial f_i/\partial u_j$, and all partial derivatives are evaluated at the operating point.

For the linear system (4), there is a well-developed and extensive theory of the system behavior as a function of its structure, expressed in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. One may broadly distinguish two parts of the theory, named classical control theory (e. g., [38]) and modern linear systems theory (e. g., [4,31]). We return to the classical control theory in the next section.

Modern control theory or linear systems theory (LST) is concerned with the dynamical properties of the system as a direct function of the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. A key element in this theory is the notion of the system *eigenvalues*, i. e., the eigenvalues of the matrix $\boldsymbol{A}$. If, for simplicity, we restrict ourselves to the endogenous dynamics of the system (set $\boldsymbol{u} = 0$), we can write the solution to (4) as

$$x_i(t) = c_{i,1}\exp(\lambda_1 t) + c_{i,2}\exp(\lambda_2 t) + \cdots + c_{i,n}\exp(\lambda_n t) , \quad i = 1, \ldots, n , \quad (5)$$

where $\lambda_1, \ldots, \lambda_n$ are the n eigenvalues of the matrix $\boldsymbol{A}$ and $c_{i.j}$ are constants that depend upon the eigenvectors and the initial condition of the system. In other words, the resulting behavior is a weighted sum of distinct *behavior modes*, $\exp(\lambda t)$. If an eigenvalue is real, the corresponding behavior mode is exponential growth (if $\lambda > 0$) or exponential decay (if $\lambda < 0$). Complex-valued eigenvalues come in complex conjugate pairs $\lambda = \tau \pm i\omega$ which give rise to oscillations $\exp(\tau\, t)\sin(\omega t + \phi)$ of frequency ω that are either expanding (if $\tau > 0$) or damped (if $\tau < 0$). In this manner, the eigenvalues serve as a compact and rigorous characterization of the behavior (of linear systems).

At any point in time, any system, linear or nonlinear, may be approximated by the expression (5). Whether it remains a good approximation depends upon how much and how quickly the eigenvalues change due to the nonlinearities in the function $\boldsymbol{f}$. If they are more or less constant for significant periods of time, we may speak of *quasilinear systems* that are well approximated by the linear system.

In some cases, however, the eigenvalues change so rapidly that it makes little sense to characterize the behavior by equation (5). (See [29] for further discussion).

Traditional Control Theory Approaches

The first set of methods, which we call the traditional approach, has been used for decades and is part of the standard curriculum in system dynamics teaching at the graduate level. It involves using the concepts from classical control theory [38] to very simple systems with only a few state variables.

The starting point is the simple first- and second-order positive and negative feedback loops found in any introductory treatment of system dynamics. The advantage of the approach is its simplicity. Although it serves at a guide to intuition, however, the obvious shortage is that it applies rigorously only to simple systems. There have been some attempts to treat higher-order systems by adding a few feedback loops [23], but the step to large-scale models is beyond this method given its inherent limitations.

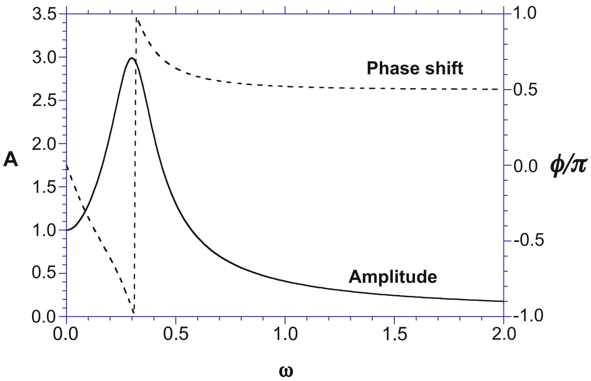
Graham [23] distills a number of principles that are based on the metaphor of a "disturbance" traveling along the chain of causal links in a feedback loop and getting amplified, damped, and possibly delayed in the process. For major negative feedback loops, which are known to tend to produce oscillation, adding minor negative loops and cross-links, or shortening the delay times increases the damping. Conversely, adding positive loops in to the oscillatory system tends to lengthen the period of oscillation whereas the effect on the damping depends upon the delays in the positive loop. Using the metaphor of pushing a child on a swing, it becomes clear that the timing of the propagation of a disturbance has as much importance for its effect on the damping as its strength.

For analyzing the behavior of positive feedback loops, Graham suggested calculating the Open-loop steady-state gain (OLSSG), a measure of the amplification around the loop. A gain greater than unity will result in exponential growth while gains less than 1 will give exponential adjustment (leveling off or decay). The intuition is perhaps best illustrated by an example: sales-driven growth. Suppose a salesperson can eventually pull in $100,000 per month in orders (probably with a several-month long delay), and assume that the company allocates 10% of revenue to marketing. Then this eventually leads to $10,000 per month for sales efforts. If the cost of a salesperson (salary, overhead, expenses etc.) is, say, $8,000 dollars per month, then the efforts of the current sales force will provide enough revenue to support $10,000/8,000 = 1.25$ persons per current person. Thus, the OLSSG of the positive loop from salespersons → orders → revenues → marketing budget → salespersons is 1.25, and the system will grow exponentially (until other factors limit the growth). Conversely, if the gain is less than 1, one salesperson will not sell enough to support their own cost, and the loop will lead to exponential decay. Graham showed how the actual rate of growth is partly determined by the OLSSG, and partly by the time constants (delays etc.) involved. (See also Subsect. 15.3 in [59]).

In the context of oscillating systems, system dynamics has also employed a concept from classical control theory, frequency response. The frequency response is determined from the transfer function of the system, $G(i\omega)$, which is a complex-valued function that specifies how an input signal $u(t)$ with frequency ω results in an output signal $x(t)$ that may be phase shifted (delayed), and either amplified or attenuated. For linear systems, G can be calculated directly from the system matrices in (4) – the transfer function (matrix) is $\boldsymbol{G}(i\omega) = \boldsymbol{B}(i\omega \boldsymbol{I} - \boldsymbol{A})^{-1}$, where $\boldsymbol{I}$ is the identity matrix (see e. g. [4]). For nonlinear systems, G may be found through simulation experiments.

Usually, G is represented in a *Bode* or *phase-and-gain* diagram. For instance, Fig. 4 shows a Bode diagram of the inventory variable INV(t) relative to the exogenous demand input variable DEM(t) in the inventory workforce model in Fig. 3. The diagram shows how the relative amplitude of the oscillation and the relative phase shift (in radians) between input and output varies as a function of the frequency of the input.

It is clear from the diagram that there is a certain frequency range, around the system's own natural frequency,



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 4
Phase-and-gain diagram (Bode diagram) showing the inventory $A\sin(\omega t + \phi)$ with amplitude A and phase shift ϕ, relative to a sinusoidal demand $\sin(\omega t)$, for varying values of the frequency ω of the demand fluctuation

where fluctuations in demand are greatly amplified compared to other frequencies. Indeed, it is a general phenomenon in systems that they will tend to amplify certain frequencies while attenuating other frequencies. This may be used to explain or understand the role of particular structures in the model in generating oscillation at certain frequencies, even when there are no oscillations coming in from the outside world. (External random noise is enough to produce oscillations in the system because random noise contains fluctuations at all frequencies). In this manner, the approach nicely demonstrates the "endogenous viewpoint" that behavior (oscillations) is generated internally by the system. As an analytic tool for large scale systems, however, the method does not seem to produce any additional insights. Thus, we may conclude that the classical approaches serve mostly as intuitive metaphors to guide the analyst rather than as full analytical tools.

Pathway Participation Metrics

The *pathway participation* method [34,35] represents a further development of an original suggestion by Richardson [46] to provide a rigorous definition of loop polarity and loop dominance. Richardson motivated this with the common confusion associated with positive feedback loops, which may exhibit a wide range of behaviors [23], as Barry Richmond noted with wonderful humor:

"Positive loops are … er, well, they give rise to exponential growth … or collapse … but only under certain conditions … Under other conditions they behave like negative feedback loops …" [49].

Richardson proposed that the polarity of a loop be defined as the sign of the expression

$$\frac{\partial \dot{x}_i}{\partial x_i} = \frac{\partial f_i\,(\boldsymbol{x}\,,\boldsymbol{u})}{\partial x_i}\,, \tag{6}$$

in the model (1), with a positive sign indicating a positive loop and vice-versa. When several loops operate simultaneously, the sign of the expression indicates whether the positive or negative loops dominate. Note, however, that the definition only applies to minor loops (i. e. loops involving a single level). Put differently, it only considers the diagonal elements of the matrix $\boldsymbol{A}$ in the linearized system (4). Richardson [46] demonstrates how even with this limitation, analyzing the system with this metric can (sometimes) yield insights into behavior of higher-order systems.

The expression (6) hints that it is relevant to consider the curvature, i. e., the second time derivative, $\ddot{x}$, of a variable when looking for dominant structure. Although he does not say so explicitly, this is effectively the focus of Mojtahedzadeh's pathway method. Figure 5 shows how one may classify behavior by comparing the first and second time derivatives of a variable. As seen in the figure, the sign of the expression $\ddot{x}/\dot{x}$, which Mojtahedzadeh denotes the total *pathway participation metric* or *PPM*, indicates whether the behavior appears dominated by positive or negative loops, much in line with Richardson's definition of dominant polarity. A zero curvature indicates a shift in loop dominance (cf. the middle column in the figure). Note, however, that the interpretation of the middle row in the figure where the slope $\dot{x}$ is zero has no clear interpretation in terms of loop dominance. Indeed this hints at one of the weaknesses of the approach that we will return to below.

Mojtahedzadeh's method proceeds by decomposing the PPM into its constituent terms as follows,

$$\mathrm{PPM}_i = \frac{\ddot{x}_i}{\dot{x}_i} = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}\frac{\dot{x}_j}{\dot{x}_i}\,, \tag{7}$$

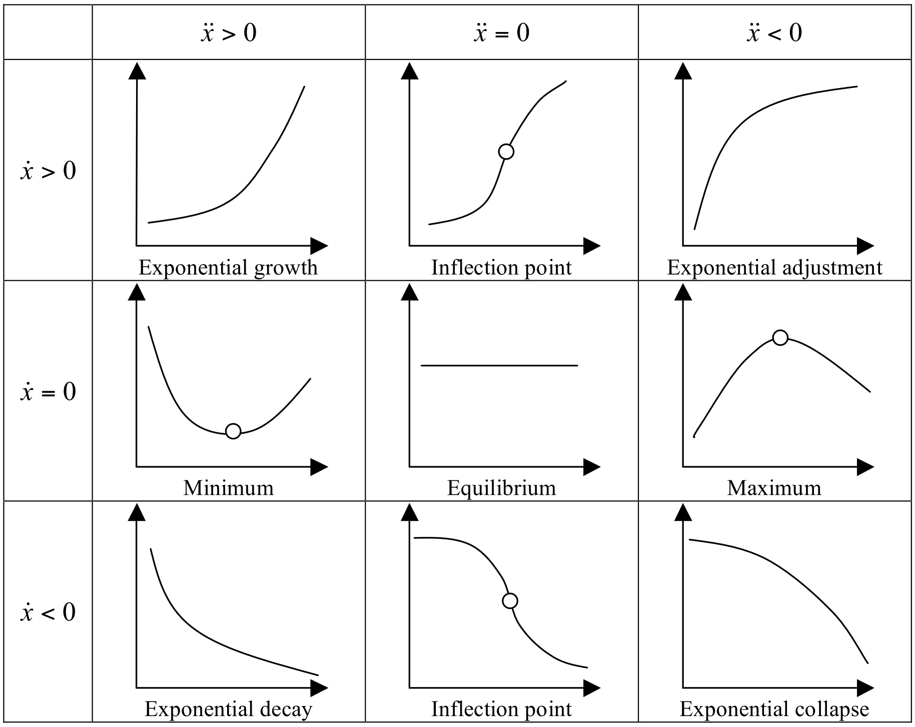
where, for brevity, we have chosen to ignore the exogenous variables u. One might say that each of the terms in the sum in (7) represents the separate influence of each of the systems' state variables on the behavior of x_i. Mojtahedzadeh in fact uses a normalized measure for the terms,

$$\frac{\left(\partial f_i/\partial x_j\right)\dot{x}_j}{\sum\limits_{k=1}^{n}\left|\left(\partial f_i/\partial x_k\right)\dot{x}_k\right|}\,, \tag{8}$$

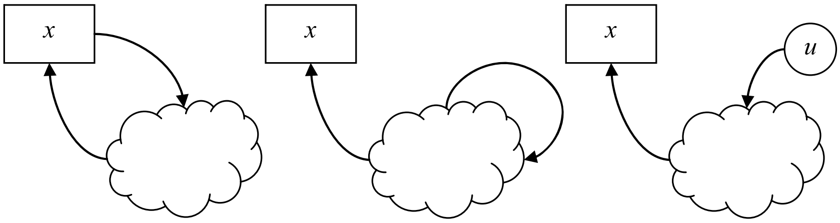
which can vary between −1 and +1, to measure the relative importance of the pathway from variable j. By explicitly considering auxiliary variables y in the model, one may further decompose each term $\partial f_i/\partial x_j$ into a sum of terms

$$\frac{\partial f_i^k}{\partial x_j} = \frac{\partial f_i}{\partial y_1}\cdot\frac{\partial y_1}{\partial y_2}\cdot\dots\cdot\frac{\partial y_{m-1}}{\partial y_m}\cdot\frac{\partial y_m}{\partial x_j}\,, \tag{9}$$

corresponding to a causal chain or pathway $\pi_k = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$. Mojtahedzadeh now considers each possible pathway (9) and defines the "dominant" pathway as the one with the largest numerical value and the same sign as PPM_i. Having selected this dominant pathway, $\pi_{ij}^* = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$, which originates in the state variable x_j the procedure is repeated for that state variable x_j, and so forth, until one either reaches one of the already "visited" state variables (in which case a loop has been found) or an exogenous variable (in which case an external driving force has been found). Thus, the procedure may result in three alternative forms of dominant structure illustrated in Fig. 6, namely



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 5
Characteristic behavior patterns based on the first and second time derivatives



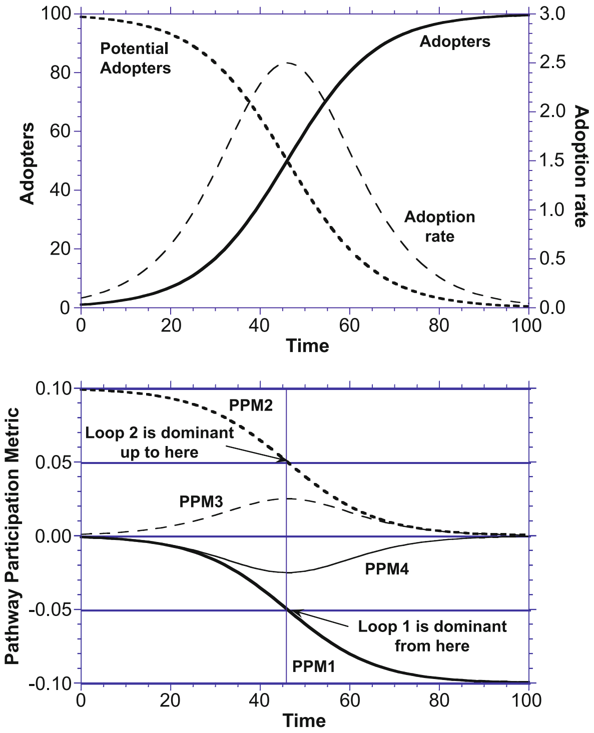
System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 6
Three alternative forms of dominant structure in the PPM method

a "pure" minor or major feedback loop, a pathway from a feedback loop elsewhere in the system, or a pathway from an exogenous variable.

By dividing the observed model behavior into different phases according to the taxonomy in Fig. 5 and then applying the method just described at different points in during these phases, one can reveal how the dominant structure changes over time. For illustration, the PPM method is applied to the Bass model and the results are presented in Fig. 7. The figure shows the metrics of four alternative pathways (four feedback loops) and the results accord nicely with the informal analysis done earlier: The method identifies two phases, exponential growth, exponential adjustment, and identifies the "word-of-mouth" positive loop (loop 1) as dominant in the first phase and the "exhaustion" loop (loop 2) as dominant in the second phase.

The PPM method is still mostly used at an early explorative stage on rather simple models, where it does appear to aid insight into the dynamics (e. g. [41]), and has been implemented in a software package, *Digest*, [35].

From the studies performed so far, it is clear that the main strength of this method is its relative computational simplicity (it does not require computing eigenvalues, which is a numerically demanding task), and the intuitive and direct connection it makes between the observed behavior and the influencing structural elements. Unlike the other approaches which operate in the "frequency do-



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 7
Pathway participation measures in the Bass model

main", the method considers the time path of a specific variable directly.

There are, however, some important outstanding issues that remain to be clarified. First, the method is not suitable for oscillatory systems. The problem is easy to recognize when one considers how the PPM measure will vary over the course of a sinusoidal outcome: The sign of the PPM will shift twice during each cycle, indicating that the behavior is alternately dominated by positive and negative loops, even though the system structure, and hence the loop dominance, may remain unchanged all the time. Richardson [46] already alluded to this problem by noting that the measure only considers the diagonal elements in the system matrix in (4), yet we know that the structure causing oscillation is the major negative loop that involve the off-diagonal elements. This is a significant limitation, given the prevalence and importance of oscillation in system dynamics analysis.

A second limitation of the current implementation of PPM is that it uses a depth-first search for the single most influential pathway for a variable. This strategy does not capture the situation where more than one structure may contribute significantly to the model behavior and, through the depth-first algorithm, may miss alternative paths that could prove to yield a larger total value of the metric. This problem could be addressed by modifying the search algorithm and is most likely of minor importance.

Another issue is how to treat the case when $\dot{x} = 0$ since it appears in the denominator of the terms in (6). However, it is not clear that it is necessary to do this division, given that it is easy to identify the nine cases in the figure by simply examining its sign. Thus, the issue is probably not of much significance.

The fourth issue, on the other hand, is more significant, namely the emphasis on identifying a single "dominant" structure. In reality, of course, the behavior of a variable is influenced by many loops and pathways at once. Reducing the consideration to a single one of these may miss important features of the structure-behavior relationships. For instance, a variable may be influenced by two negative loops and one positive, with the sum of the two negative loops dominating the influence of the positive loop, even though that loop by itself has the strongest influence on the behavior. It is more appropriate to consider the relative importance of alternative pathways, yet the method does not address how one would partition the behavior among pathways (the three structures in Fig. 6) – only among individual links.

Thus, while the notion of pathways seems an interesting and useful idea, it may be that it will ultimately be more effective to use a list, ranked in order of magnitude, of the pathways that influence a variable.

Finally, the method shares a weakness with the traditional method in that it considers primarily partial system structures rather than global system properties. In contrast, the two eigenvalue methods to which we now turn are based on a rigorous characterization of the entire system (at a given point in time).

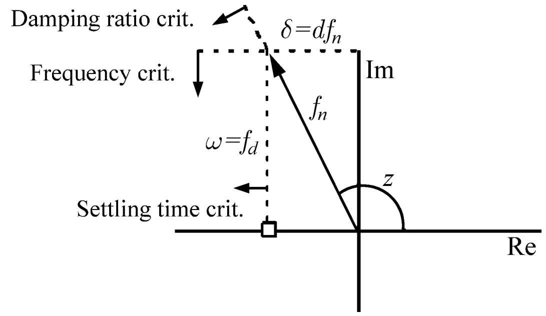
Eigenvalue Elasticity Analysis

The third method may be termed *eigenvalue elasticity analysis* (or EEA for short) and builds upon the tools from modern linear systems theory (LST), applied to the linearized model (4). The method is concerned with the structural elements that significantly affect the system eigenvalues or behavior modes – the values λ in (5). Specifically, it measures influence by the elasticity of an eigenvalue λ with respect to some parameter g in the model, defined as $\varepsilon = \left(\partial\lambda/\partial g\right)(g/\lambda)$, i. e. the fractional change in the eigenvalue relative to the fractional change in the parameter. The advantage of this fractional measure is that it is dimensionless, i. e., independent upon the choice of units, including the time scale unit. Sometimes, the influence measure is used instead, defined as

$\mu = (\partial\lambda/\partial g)g$. This measure has dimension $[1/time]$ and so depends upon the choice of is time unit, but it is generally easier to interpret for complex-valued eigenvalues and avoids numerical problems with very small or zero eigenvalues (see [29,54]).

The idea behind EEA was first introduced in system dynamics by Forrester [14] in the context of economic stabilization policy. For purposes of policy analysis in oscillating systems, one may define a number of criteria from engineering control theory, all of which relate to the eigenvalues of the system, as summarized in Table 2. Figure 8 provides a graphical characterization of the eigenvalues and policy criteria in the complex plane. Though these measures are not new, the EEA method is unique in its attempt to use them to gain qualitative intuitive understanding of the system. A significant step in this direction was first suggested by Forrester [15] with the notion that the elasticities of any links in the model (corresponding to elements of the matrix $\boldsymbol{A}$ in the linearized system (4)), can be interpreted as the sum of elasticities of all feedback loops containing that link. We have chosen to name this approach *loop eigenvalue elasticity analysis (LEEA)*.

Kampmann [28] provided a rigorous definition of LEEA and also pointed to the fact that feedback loops are not independent. In other words, given the possibly very large number of loops in a given model (Kampmann demonstrated how the theoretical maximum number of loops grows combinatorically with the number of variables), it only makes sense to speak of individual contributions of a limited set of *independent* loops. He proved that a fully connected system (where there is a feedback loop between any pair of variables – the typical case in system dynamics models) with N links and n variables has a total of $N - n + 1$ independent loops and provided a procedure for constructing this set and calculating the loop elasticities.



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 8
Characterization of eigenvalues plotted in the complex plane

Kampmann's analysis points to a fundamental issue relating to the notion of feedback loops as a way to explain behavior: the significance assigned to a particular loop depends upon the context (the chosen independent loop set). In other words, feedback loops are derived and relative concepts rather than fundamental independent building blocks of systems. Oliva [40] further refined the definition of independent loop sets by introducing the *Shortest independent loop set (SILS)* along with a procedure for constructing the set. Although a SILS is not generally unique, experience seems to suggest that it is easier to interpret [41]. Yet the issue remains that independent feedback loop sets are relative concepts.

System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Table 2
Stabilization policy criteria and corresponding effects on eigenvalues and BDW of a policy change in a system element *g*

Policy Criterion	Description	Change in eigenvalue $\lambda = \delta \pm i\omega, \omega > 0$	Change in BDW $\boldsymbol{w}$	Appropriate measure in time path
Damping	Increases the rate of decay of oscillation (or decreases the rate of expansion)	$\frac{\partial\delta}{\partial g}\frac{g}{\delta} < 0$	N/A	$\frac{x(t+T)}{x(t)}$
Frequency	Decreases the frequency of oscillation (or lengthens the period T)	$\frac{\partial\omega}{\partial g}\frac{g}{\omega} < 0$	N/A	T
Variance	Reduces the variance of a target variable (or the weighted average variances of several variables)	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	$\int x(t)^2\, dt$
Auto-spectrum	Reduces variance of target variable(s) within a target frequency range	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	Filter in frequency domain
Frequency response gain	Reduces the gain (amplification) in the target frequency range for a particular combination of disturbance exogenous and output variables.	Based upon transfer function $G(i\omega)$		

System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Table 3
Loops and their influences in the inventory workforce model. Values are measured at time $t = 0$. The model contains three eigenvalues, $\lambda_1 = -0.250$ and $\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$. The influence measure is defined as $g \cdot \partial\lambda/\partial g$. For the imaginary part, a positive influence measure means that the frequency is increased

Loop	Nodes	Gain	Influence on Re[λ_1]	Influence on Re[λ_2]	Influence on Im[λ_2]
1	ED > CED	−0.250	−0.250	0.000	0.000
2	W > HFR	−0.200	0.000	−0.100	−0.022
3	INV > IC > DP > SP > EO > P	−0.076	0.000	−0.038	0.008
4	INV > IC > DP > DW > HFR > W > NP > P	−0.100	0.000	0.000	0.176
5	INV > IC > DP > DW > HFR > W > NP > SP > EO > P	0.015	0.000	0.000	−0.027

In Table 3, we show how the LEEA analysis applies to the simple inventory–workforce model in Fig. 3. The model contains a total of 5 feedback loops, all of which are independent. The loops are listed in Table 3, including their constituent variables (nodes), and the gain of the loop (defined in a similar manner to the pathway participation metrics above). We see that there are three minor negative loops, related to the exponential smoothing of expected demand (loop 1) and the adjustment of workforce to desired workforce (loop 2). The minor loop 3 is the "overtime shortcut" that allows production to adjust part way to desired production immediately so one does not have to wait for the workforce to adjust. Loop 4 is the main major negative loop that adjusts inventory to desired levels via workforce adjustment. Finally, loop 5 (the only positive loop) is a fairly weak loop that moderates the effect of loop 4 by adjusting the overtime effect "back to normal" when the workforce is brought in line with desired production.

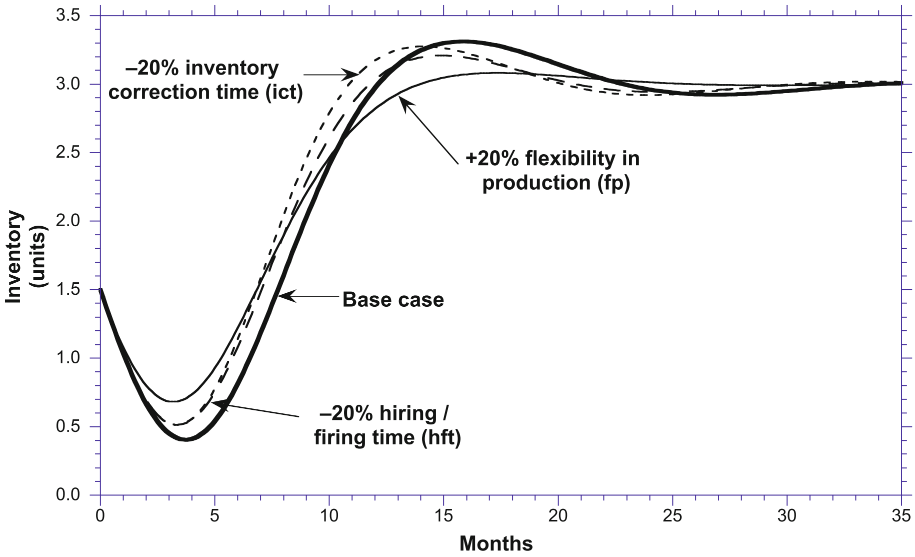
Although the model is nonlinear (due to the overtime function), the eigenvalues do not change very much over the course of its behavior. The model contains one real eigenvalue ($\lambda_1 = -0.250$) and one pair of complex conjugate eigenvalues ($\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$). The first eigenvalue corresponds to the adjustment of expected demand (ED). The other pair produces a damped oscillation in inventory and workforce.

Table 3 also shows the loop influences upon the three eigenvalues. Note how there is a one-to-one correspondence between loop 1 (the adjustment of expected demand) and the first eigenvalue. This is due to fact that the ED level constitutes a single strongly connected component of the model (see Fig. 3), i. e. there is no feedback between this level and the rest of the model. We also note that the workforce adjustment and the overtime loops have a stabilizing influence upon the behavior (they make the real part of the oscillatory eigenvalues more negative and have relative little effect upon the frequency of oscillation). Conversely, the major negative loop 4 has a destabilizing influence, since strengthening it will increase the frequency of oscillation and not increase the damping. The effects of loop 5 are fairly weak.

From this analysis, one would therefore expect parameters that strengthen loop 2 (shortening hire/fire time) or loop 3 (increase overtime effect) would stabilize the system while strengthening loop 4 (shorter inventory adjustment time) will destabilize the system. Indeed this is what happens, as illustrated in the simulations in Fig. 9.

The EEA/LEEA method has been applied in a number of contexts (e. g. [1,20,22,24,29,51,52,54]), but remains a tool employed only by specialists in fundamental research, not least because it has not been incorporated into standard software packages. Thus, the potential of the method for widespread practice remains unexplored.

One might be skeptical that a method derived from linear systems theory may have any use for the nonlinear models found in system dynamics. Kampmann and Oliva [29] considered what types of models the method would be particularly suited for. They defined three categories of models, based upon the behavior they are designed to exhibit: 1) linear and quasilinear models, 2) nonlinear single-transient models, and 3) nonlinear periodic models. The first category encompasses models of oscillations, possibly combined with growth trends, with relatively stable equilibrium points, (e. g., the classical industrial dynamics models [11]). Nonlinearities may modify behavior (particularly responses to extreme shocks) but the instabilities and growth trends can be analyzed in terms of linear relationships. Kampmann and Oliva concluded that LEEA showed the most promise and potential for this class of models because the analytical foundations are solid and valid, and because the method has the ability to find high-elasticity loops even in large models very quickly without much intervention on the part of the analyst.



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 9
Simulated behavior of inventory–workforce model, showing the effect on inventory of parameter changes for overtime (flexibility of production), inventory adjustment time (ict), and labor hiring/firing time (hft), respectively

The second class is typical of scenario models like the World Model [13,33], the Urban Dynamics Model [12], or the energy transition model in [57], to name a few, that show a single transient behavior pattern, like overshoot and collapse or a turbulent transition to a new equilibrium. In these models, nonlinearities usually play an essential role in the dynamics. Yet it is possible to divide the behavior into distinct phases where certain loops tend to dominate the behavior. In this class of models LEEA also shows promise by measuring shifts in structural dominance by the change in elasticities. But it requires more input from the analyst (e. g. in defining the different phases of the transition) and it has no obvious advantage over other methods, like PPM.

The third class, nonlinear periodic models, are those that exhibit fluctuating behavior in which nonlinearities play an essential role, such as like limit cycles, quasiperiodic behavior, or chaos, (see, e. g. [48]). Here the utility of the method is much less clear and depends upon the specifics of the model in question. For example, the classic Lorenz model that exhibits limit cycles, period doubling and deterministic chaos does not lend itself to any insight using LEEA [29]. This is particularly the case in systems with strong nonlinearities such as min and max functions. In these systems, the behavior may change abruptly (eigenvalues suddenly shift) in what is called border-collision bifurcations [37,61]. In other cases, the method of breaking the behavior into phases with different dominant structures may yield significant insight from LEEA. For instance, Sterman's simple long wave model [58] lends itself well to this approach (e. g. [24,28]).

In the present paper, we add a fourth category of models or behavior for which the method has not been explored yet. We name this category nonlinear multi-modal models. These encompass the cases where one behavior mode interacts with and therefore modifies another behavior mode – something that can only happen in nonlinear systems. The most common example is mode-locking or entrainment, in which oscillations become synchronized (e. g. [25]). Another example is mode modification, where one behavior mode (growth or oscillation) affects the character of another (typically oscillation). An example of this is the interaction of the business cycle with the economic long wave, where the former tends to get more severe during long wave downturns [16]. Whether LEEA can contribute to this class of models remains to be seen.

Compared to the former two methods, the EEA/LEEA is mathematically more general and rigorous, though many of the mathematical issues in the method remain to be addressed, as we summarize below. This rigor is also the main strength of the method, since it provides an unambiguous and complete measure of the influence of the entire feedback structure on all behavior modes.

A weakness or challenge that is starting to show up is the computational intensity in calculating eigenvalues and elasticities. This is not so much an issue of computer time and memory space as of the stability of numerical methods. Kampmann and Oliva [29] found that the numerical method used sometimes proved unstable, yielding meaningless results. Clearly, there is a need to explore this issue further, possibly building upon the developments in control engineering.

A more serious weakness is the difficulty in interpreting the results: Eigenvalues do not directly relate to the observed behavior of a particular variable. The concepts of eigenvalues and elasticities are rather abstract and unintuitive [10]. There is a need for tools and methods that can translate them into visible, visceral, and salient measures. Here, the measures in Table 2 may provide a guide. In particular, it is possible to use (linear) filtering in the frequency domain to define a behavior of interest. For example, an analyst may be concerned with structures causing a typical business cycle (3–4-year oscillation) and, by specifying a filter that "picks out" that range of fluctuation, could obtain measures for structures that have elasticities in that range. Because filters are typically linear operators, all the analytical machinery of the LEEA method will also apply in this case – a significant advantage.

Using filters will also solve an issue that appears in large-scale models, namely the presence of several identical or nearly identical behavior modes. Saleh et al. [54] do consider the analytical problems associated with repeated eigenvalues, where it becomes necessary to use generalized eigenvectors, and where other behavior modes appear involving power functions of time. A filter essentially constitutes a weighted average of behavior modes and in this fashion avoids the "identity problem" of non-distinct eigenvalues.

The most serious theoretical issue, in our view, is how the results are interpreted using the feedback loop concept. As mentioned, the concept is relative (to a choice of an independent loop set). Moreover, practice reveals that the number of loops to consider is rather large and that the loops elasticities often do not have an easy or intuitive explanation. A lot of care must be taken when interpreting the results. For instance, Kampmann and Oliva [29] found that "phantom loops" – loops that cancel each other by logical necessity and are essentially artifacts of the equation formulations used in the model – could nonetheless have large elasticities and thus seriously distort the interpretation of the results. An example of "phantom loops" is found in the Bass model in Fig. 1, where loops 3 and 4 are artifacts of the way the model is formulated. If the variable *Total population* (T) was eliminated from the equations, the loops would disappear and in fact they exactly cancel each other out (since T is constant). Nonetheless, they appear on the list of loops and appear to have a separate influence on behavior. These kinds of problems may not be intractable, but their resolution will require careful mathematical analysis.

Finally, a problem with EEA and LEEA is that it only considers changes to behavior modes, not the degree to which these modes are expressed in a system variable of interest. This issue is addressed by also considering the eigen*vectors* of the system, which is the foundation for the analysis in the next section.

Eigenvectors and Dynamic Decomposition Weights (DDW)

The last set of methods, which are still in early development, we have termed the *eigenvector-based* approach (EVA). EVA attempts to improve the EEA/LEEA method by considering how much an eigenvalue or behavior mode is expressed in a particular system variable. The logic of the method and how EEA and EVA complement each other is shown in Fig. 10. As shown by Kampmann [28], in a sense there is a one-to-one correspondence between eigenvalues and loop gains whereas the eigenvectors arise from the remaining "degrees of freedom" in the system. The observed behavior of the state variables in the model is then the combined outcome of the behavior modes (from the loop gains) and the weights for each mode (from the eigenvectors) in the respective state variable.

A number of researchers have attempted to develop EVA methods. Some emphasize the curvature (second time derivative) of the behavior, similar to the starting point of the PPM method [24,50,51,52]. The slope or rate of change $\dot{x}(t)$ of a given variable x in the linearized system may be written by

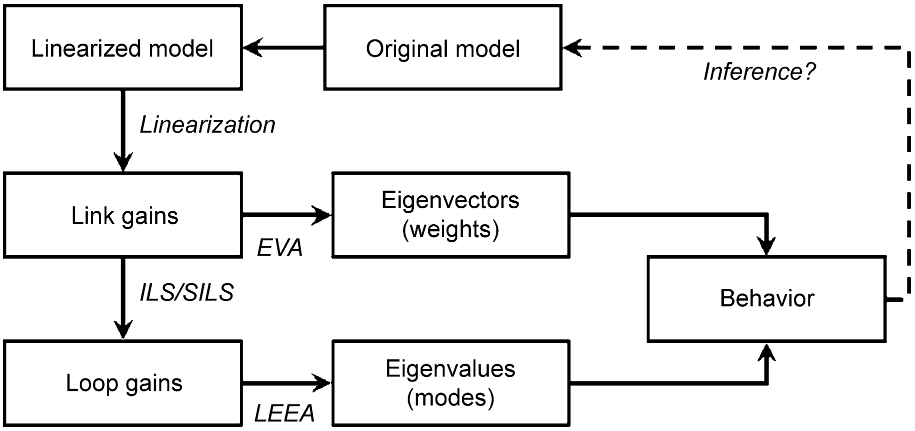
$$\dot{x}(t-t_0) = w_1 \exp(\lambda_1(t-t_0)) + \cdots + w_n \exp(\lambda_n(t-t_0)) , \quad (10)$$

where the weights w_i are related to the eigenvectors. Then the curvature at time t_0 is

$$\ddot{x}(t_0) = w_1\lambda_1 + \cdots + w_n\lambda_n . \quad (11)$$

One may therefore interpret (11) has the sum of contribution from individual behavior modes. Güneralp [24] suggested using the terms on the right-hand side of (11) as weights to combine elasticities of individual behavior modes ε_i with respect to some system element (like a link gain or a loop gain) into a weighted sum

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} w_i\lambda_i\varepsilon_i}{\sum_{i=1}^{n} |w_i\lambda_i|} , \quad (12)$$



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 10
Schematic view of eigenvalue and eigenvector analysis approach

as a measure of the overall significance of that system element. He further normalized the elasticity measure by the elasticity measure for other system elements, i. e., assuming there are K such elements (loops or links), the relative importance ρ_k of the kth element is defined as

$$\rho_k = \frac{\bar{\varepsilon}_k}{\sum_{j=1}^{K} \left|\bar{\varepsilon}_j\right|} , \tag{13}$$

with the motivation that elasticities may vary greatly in numerical values, making comparisons at different points in time difficult, whereas ρ_k is a relative measure varying between +1 and −1. His results shed an alternative light on the behavior of these models, but the mathematical meaning, consistency and significance of the doubly normalized measure (13) remains to be clarified. It is still too early to tell what the most useful approach will be, but one may note that the emphasis on the curvature shares the basic weakness in the PPM approach in dealing with oscillations.

Other researchers have looked directly at the *dynamic decomposition weights (DDW)* w_i in (10), i. e., the relative weight of the modes for a particular variable, from a policy criterion perspective, similar to Forrester's original focus and the starting point for the EEA analysis [21,53,54].

For instance, Saleh et al. [54] look at how alternative stabilization policies affect the behavior of business cycle models, using both a simple inventory–workforce model [59], and a more extensive model based on Mass [32] and used in the LEEA analysis of Kampmann and Oliva [29]. Using the procedure in Fig. 9, they decompose the net stabilizing effect of a policy into its effect on the behavior mode itself (LEEA) and its effect on the expression of that mode in the variable of interest, measured the dynamic decomposition weights (EVA or DDW).

To illustrate the approach we perform the computations for the inventory–workforce model (Fig. 3 and Table 1). We find that the following equations describe the behavior of the state variables

$$\begin{aligned} \text{ED} &= 1 - 0.500\mathrm{e}^{-0.250t} \\ \text{INV} &= 3 - 2.167\mathrm{e}^{-0.250t} \\ &\quad + 1.134\mathrm{e}^{-0.138t}\sin(2.945 + 0.285t) \\ \text{WF} &= 1 + 0.669\mathrm{e}^{-0.250t} \\ &\quad - 1.169\mathrm{e}^{-0.138t}\sin(1.553 - 0.285t) \ . \end{aligned} \tag{14}$$

As expected from the structure of the model, the behavior of *Expected Demand* does not have an oscillatory component and only shows a short transient exponential adjustment for the stock to match *Demand*. On the other hand, *Inventory* and *Workforce*, in addition to having the transient behavior to reach equilibrium captured by the first eigenvalue, have an oscillatory component represented by the second eigenvalue. Note that each state variable has a different *Dynamic Decomposition Weight* (w) for each reference mode, i. e., each eigenvalue contributes differently to the overall behavior of each state variable.

An exploration of the policy design space can be achieved by assessing the influence of model parameters on the dynamic decomposition weight. By focusing on the weights of the behavior modes for the variable of interest we can identify leverage points to increase or decrease the presence of a behavior mode in the variable. The weight elasticity column in Table 4 reports the parameter elasticity of w_2 (the weight of eigenvalue 2, the oscillatory behavior mode) on *Inventory* ($\varepsilon_w = (\mathrm{d}w/\mathrm{d}p)(p/w)$). The mag-

System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Table 4
Elasticity to parameters of weight of eigenvalue 2 ($-0.138 + i\,0.85$) on inventory and influence of parameters on eigenvalue 2 – inventory–workforce model

Parameter	w_2 on INV Elasticity	Influence on Re[$\boldsymbol{\lambda}_2$]	Influence on Im[$\boldsymbol{\lambda}_2$]
Demand	2.000	0.000	0.000
Inventory correction time	0.656	0.038	−0.157
Flexibility in production	0.364	−0.549	−0.266
Productivity	−0.353	0.000	0.000
Time to change expectations	−0.240	0.000	0.000
Normal inventory coverage	0.239	0.000	0.000
Hiring/Firing time	0.238	0.100	−0.127
Disequilibrium fraction	0.000	0.000	0.000

nitude of the elasticity quantifies the impact that changes in the parameter value have on the weight of the oscillatory behavior model on *Inventory*. The table is sorted in descending order of absolute value of elasticity.

Changes in parameters, however, not only impact the behavior decomposition weights, but also change the eigenvalues themselves. This dual impact of parameter changes introduces a challenge in developing policy recommendations. The last two columns of Table 4 report the influence on the eigenvalue (real and imaginary part) for each parameter. These measures of influence should be interpreted in a similar way as the weight elasticities. The influence measure is defined as $\mu_\lambda = \left(\partial\lambda/\partial p\right) p$. A positive real-part measure indicates that increasing the parameter will destabilize the system by lengthening the settling time and vice-versa. A positive imaginary-part measure indicates that increasing the parameter will increase the frequency of oscillation – normally considered a destabilizing influence – and vice-versa.

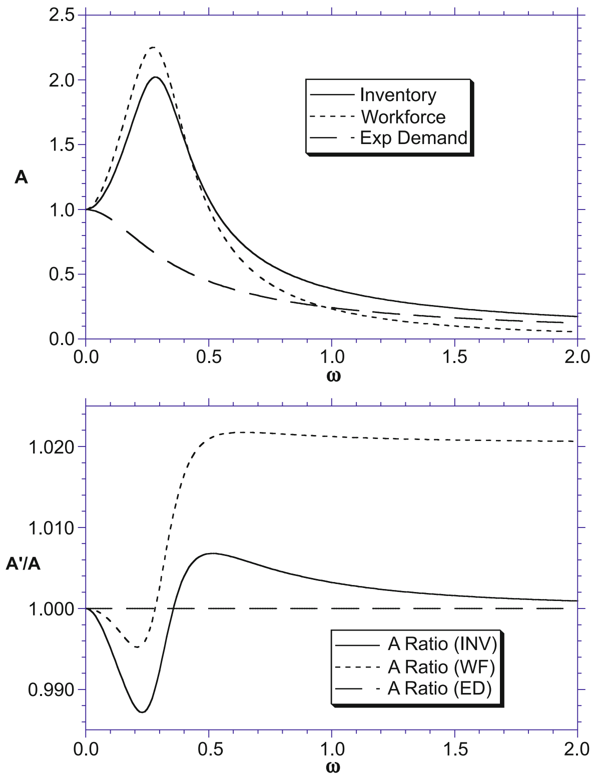
Five parameters, *demand, disequilibrium fraction*, *productivity, normal inventory coverage*, and *time to change expectations*, have no influence on the oscillatory behavior mode. *Demand* and *disequilibrium fraction* are initialization constants that do not participate in any of the feedback loops in the model. *Productivity* is essentially a scaling measure having to do with the definition of units of labor and goods in the model. Redefining units should not affect the dynamics of the model. While *time to change expectations* is involved in loop 1, it does not participate in the oscillatory behavior observed in the model since, as discuss above, *Expected Demand* is in a separate strongly connected component of the model.

In accordance with LEEA, the *flexibility in production* parameter, which strengthens overtime loop 3 (cf. Fig. 3), has a strong stabilizing influence, by both increasing the damping and lowering the frequency of the oscillatory mode. Likewise, as predicted by LEEA, a shorter *hiring/firing time* will increase damping by strengthening the labor adjustment loop 5 but, again in accordance with LEEA, also increases the frequency of adjustment because it also strengthens the major loop 4. Finally, lowering the *inventory correction time* will strengthen the link from *inventory* to *desired production*, and consequently the three loops 3, 4 and 5, with the net effect that although the adjustment is a little faster (a more negative real part), the frequency is also increased significantly, i. e., it is a less effective way of stabilizing the system (cf. Fig. 9).

As an alternative approach, Fig. 11 shows what happens to the frequency response of the state variables (*Inventory* INV, *Workforce* WF, and *Expected Demand* ED) when the parameter Hiring/Firing Time (hft) is reduced by 2% from 5 to 4.9. There are a number of things to notice in the figure. First, there is no effect whatsoever on the ED variable, which should not be surprising, given that there is no feedback to this variable from the rest of the system. Second, the effect on the amplitude, like the amplitude itself, is strongly dependent upon the frequency of variation. We see that there is a significant amount of dampening on the *Inventory* fluctuation around the resonant frequencies in the range 0.1 to 0.3. On the other hand, there is a small amount of amplification of inventory in the higher frequency ranges. The effect on *Workforce* is very different: though there is a small attenuation in the resonant frequencies, there is a significant increase in variance in the higher frequency range. In other words, although the LEEA analysis showed a faster hiring policy to be stabilizing (by strengthening loop 2, cf. Table 3), the DDW analysis shows that it depends – both upon the variable in question and the context (frequency of variation).

Future Directions

As mentioned above, it is not possible to construct a complete theory that will automatically provide modelers with



System Dynamics, Analytical Methods for Structural Dominance, Analysis in, Figure 11
Effect on frequency response of the inventory workforce model of reducing the parameter Hiring/Firing Time (hft) from 5.0 to 4.9. The diagram shows the gain of the base case (*upper graph*) for the three state variables, and the resulting change in the gain, measured as the ratio (A'/A) from the parameter change (*lower graph*)

"the" dominant structure. Given the analytical intractability of nonlinear high-order systems found in our field, the most we can hope for is a set of tools that will guide the analysis and aid the development of the modeler's intuition.

That said, however, we are left with an impression that the analytical foundation for these tools is in need of further development before one rushes into implementing them into software packages. We are quite satisfied with the current state of affairs in this regard, where code, models, and documentation are made freely to download (most of the cited papers provide a URL to their code and models). Understanding *how* and *why* the tools work the way they do is crucial, and this will require that a number of puzzles, uncertainties, and technical problems be addressed. Only then will the time come to submit the methods for wider application to test their real-world utility.

While the classical method remains a useful intuitive guide and teaching tool for graduate students, there are no signs that it may be developed further. (That said, it is possible that the classical control transfer function method may be employed in the eigensystem approaches to explore nested canonical systems, though this is purely speculative). The pathway method would benefit from a firmer mathematical foundation. In particular, it would be important to compare how its results and conclusions compare to those found in the LST. It is possible that the pathway method may eventually be merged with the LST approaches as a subset of a general analytical toolbox. We believe that there is a great deal of promise in combining the eigenvalue and eigenvector analysis in the LST approaches. This combination will yield a complete system characterization and an understanding of both how particular feedback loops are involved in generating a behavior mode, and how system elements determine the expression of that behavior mode in a particular variable. A unified LST approach along the lines suggested in Fig. 10 thus seems within reach.

It will probably be a while, however, before these methods will find their way into widely available and use-friendly software packages. Apart from the theoretical issues alluded to above, a number of technical issues related to numerical calculations, various "pathological cases" (such as non-distinct eigenvalues), and special cases of feedback loops ("figure-eight" loops, for instance), will need to be addressed.

On the more creative side, it would be interesting to explore alternative forms of visualizing the various influence measures developed. For instance, one could imagine that links between variables in a model diagram "glow" in different colors and intensities depending upon their effect on a behavior pattern in question. This is not just a question of fancy user interfaces: as mentioned in the introduction, the function of these tools will be as intuitive consistent aids to understanding, not analytical "answering machines". In this light, the visualization is as important as the analytical principles behind it. Given the power of the human eye in finding patterns in visual data, this could be a significant next step.

Bibliography

1. Abdel-Gawad A, Abdel-Aleem B, Saleh M, Davidsen P (2005) Identifying dominant behavior patterns, links and loops: Automated eigenvalue analysis of system dynamics models. Proceedings of the Int System Dynamics Conference, Boston, July 2005. System Dynamics Society, Albany
2. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Oper Res 42(1):59–87

3. Bass FM (1969) A new product growth for model consumer durables. Manag Sci 15(5):215–227
4. Chen CT (1970) Introduction to Linear System Theory. Holt, Rinnehart and Winston, New York
5. Eberlein R (1984) Simplifying Models by Retaining Selected Behavior Modes. Ph D Thesis, Sloan School of Management, MIT, Cambridge
6. Eberlein RL (1986) Full Feedback Parameter Estimation. In: Proceedings of the Int Systems Dynamics Conference, Sevilla, Spain. System Dynamics Society, Albany, pp 69–83
7. Eberlein RL (1989) Simplification and understanding of models. Syst Dyn Rev 5(1):51–68
8. Eberlein RL, Wang Q (1985) Statistical Estimation and System Dynamics Models. Proceedings of the Int Systems Dynamics Conference, Keystone, USA. System Dynamics Society, Albany, pp 206–222
9. Ford A, Flynn H (2005) Statistical screening of system dynamics models. Syst Dyn Rev 21(4):273–303
10. Ford DN (1999) A Behavioral Approach to Feedback Loop Dominance Analysis. Syst Dyn Rev 15(1):3–36
11. Forrester JW (1961) Industrial Dynamics. Productivity Press, Cambridge
12. Forrester JW (1969) Urban Dynamics. Productivity Press, Cambridge
13. Forrester JW (1971) World Dynamics. Productivity Press, Cambridge
14. Forrester N (1982) A Dynamic Synthesis of Basic Macroeconomic Policy: Implications for Stabilization Policy Analysis. Ph D Thesis, Sloan School of Management, MIT, Cambridge
15. Forrester N (1983) Eigenvalue analysis of dominant feedback loops. Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA. System Dynamics Society, Albany
16. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policymaking. Kluwer, Norwell, pp 199–240
17. Forrester JW, Senge PM (1980) Tests for Building Confidence in System Dynamics Models. TIMS Stud Manag Sci 14:209–228
18. Forrester JW, Mass NJ, Ryan CJ (1976) The System Dynamics National Model: Understanding Socio-Economic Behavior and Policy Alternatives. Technol Forecast Soc Chang 9(1–2):51–68
19. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
20. Gonçalves P (2003) Demand bubbles and phantom orders in supply chains. Ph D Thesis, Sloan School of Management, MIT, Cambridge
21. Gonçalves P (2008) Behavior modes, pathways and overall trajectories: Eigenvalue and eigenvector analysis in system dynamics. Syst Dyn Rev, forthcoming
22. Gonçalves P, Lerpattarapong C, Hines JH (2000) Implementing formal model analysis. Proceedings of the Int System Dynamics Conference, Bergen, Norway, August 2000. System Dynamics Society, Albany
23. Graham AK (1977) Principles on the Relationship Between Structure and Behavior of Dynamic Systems. Ph D Thesis, Sloan School of Management, MIT, Cambridge
24. Güneralp B (2006) Towards Coherent Loop Dominance Analysis: Progress in Eigenvalue Elasticity Analysis. Syst Dyn Rev 22(3):263–289
25. Haxholdt C, Kampmann CE, Mosekilde E, Sterman JD (1995) Mode Locking and Entrainment of Endogenous Economic Cycles. Syst Dyn Rev 11(3):177–198
26. Holland JH (1992) Adaptation in Natural and Artificial Systems. MIT Press, Cambridge
27. Homer JB (1983) Partial-Model Testing as a Validation Tool for System Dynamics. In: Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA, July 1983. System Dynamics Society, Albany, pp 920–932
28. Kampmann CE (1996) Feedback Loop Gains and System Behavior (unpublished manuscript). In: Proceedings of the Int System Dynamics Conference, Cambridge, USA, July 1996. System Dynamics Society, Albany, pp 260–263
29. Kampmann CE, Oliva R (2006) Loop Eigenvalue Elasticity Analysis: Three Case Studies. Syst Dyn Rev 22(2):146–162
30. Lane DC, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12(2):87–120
31. Luenberger DG (1979) Introduction to Dynamic Systems: Theory, Models and Applications. Wiley, New York
32. Mass NJ (1975) Economic Cycles: An Analysis of Underlying Causes. Productivity Press, Cambridge
33. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
34. Mojtahedzadeh MT (1996) A path taken: Computer-assisted heuristics for understanding dynamic systems. Ph D Thesis, Rockefeller College of Pubic Affairs and Policy, State University of New York at Albany, Albany
35. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
36. Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916
37. Mosekilde E, Laugesen JL (2006) Nonlinear Dynamic Phenomena in the BEER Model. Department of Physics, The Technical University of Denmark, Kongens Lyngby
38. Ogata K (1990) Modern Control Engineering, 2nd edn. Prentice Hall, Englewood Cliffs
39. Oliva R (2003) Model Calibration as a Testing Strategy for System Dynamics Models. Eur J Operat Res 151(3):552–568
40. Oliva R (2004) Model Structure Analysis Through Graph Theory: Partition Heuristics and Feedback Structure Decomposition. Syst Dyn Rev 20(4):313–336
41. Oliva R, Mojtahedzadeh M (2004) Keep it simple: Dominance assessment of short feedback loops. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
42. Ott E (1993) Chaos in Dynamical Systems. Cambridge University Press, New York
43. Peterson DW (1980) Statistical Tools for System Dynamics. In: Randers J (ed) Elements of the System Dynamics Method. Productivity Press, Cambridge, pp 224–241
44. Peterson DW, Eberlein RL (1994) Reality Checks: A Bridge Between Systems Thinking and System Dynamics. Syst Dyn Rev 10(2/3):159–174
45. Radzicki MJ (2004) Expectation Formation and Parameter Estimation in Uncertain Dynamical Systems: The System Dynamics Approach to Post Keynesian-Institutional Economics. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
46. Richardson GP (1984/1995) Loop Polarity, Loop Dominance, and the Concept of Dominant Polarity. Syst Dyn Rev 11(1):67–88

47. Richardson GP (1986) Dominant structure. Syst Dyn Rev 2(1):68–75
48. Richardson GP (ed) (1988) System Dynamics Review. Chaos Special Issue 4:1–2
49. Richmond B (1980) A new look at an old friend. Plexus, Resource Policy Center, Thayer School of Engineering, Dartmouth College, Hanover
50. Saleh M (2002) The characterization of model behavior and its causal foundation. Ph D Thesis, Dept of Information Science, University of Bergen, Bergen
51. Saleh M, Davidsen P (2001) The origins of behavior patterns. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
52. Saleh M, Davidsen P (2001) The origins of business cycles. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
53. Saleh M, Oliva R, Davidsen P, Kampmann CE (2006) Eigenvalue Analysis of System Dynamics Models: Another Perspective. Proceedings of the Int System Dynamics Conference, Neijmegen, The Netherlands, July 2006. System Dynamics Society, Albany
54. Saleh M, Oliva R, Davidsen P, Kampmann CE (2008) A comprehensive analytical approach for policy analysis of system dynamics models. Mays Business School, Texas A&M University, Working Paper
55. Schweppe F (1973) Uncertain Dynamical Systems. Prentice-Hall, Englewood Cliffs
56. Senge PM (1990) The Fifth Discipline: The Art & Practice of the Learning Organization. Doubleday Currency, New York
57. Sterman JD (1981) The Energy Transition and the Economy: A System Dynamics Approach. Ph D Thesis, Sloan School of Management, MIT, Cambridge
58. Sterman JD (1985) A Behavioral Model of the Economic Long Wave. J Econ Behav Org 6(1):17–53
59. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
60. Wolstenholme E (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20(4):341–356
61. Zhusubaliyev ZT, Mosekilde E (2003) Bifurcation and Chaos in Piecewise-Smooth Dynamical Systems. World Scientific, Singapore

System Dynamics, Introduction to

Brian Dangerfield
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

When Jay Wright Forrester published his first paper in 1958 he subtitled it *"a major breakthrough for decision-makers"*. At the time some thought this rather an exaggeration if not pompous. Now that 50 years of system dynamics (SD) has elapsed we can at least point to the achievements made and re-state continuing progress in the pages of this section. Was it a 'major breakthrough'? It certainly has the potential to raise the standards in evidence-based policy making to warrant this description and some startlingly good examples of such work will be mentioned here. But after 50 years perhaps one might expect more than has surfaced heretofore.

The key might be connected to the skills required to formulate good SD models – those which address a real-world problem with devastating simplicity and insight. It is deceptively easy to produce an SD model but there are subtleties involved in producing a really effective model for policy purposes. An uplift in modeling skills is something which a subset of the (now significant) amount of published material on SD is aimed at and this section will add to that corpus of work. In addition it will illustrate the extent to which SD applications have spread from its genesis in business to embrace health care, environmental, energy and climate issues, project management, some aspects of biological science and human physiology, governmental and public policy generally, economics (mainly macro), the diffusion of innovations and finally social and economic development. Other applications are being encountered as the power of the methodology is becoming appreciated. It has long since justified the change of title from ***Industrial*** Dynamics (1958) to ***System*** Dynamics (1970 onwards).

Richardson contributes an overview of the basics of SD modeling (see ► System Dynamics, The Basic Elements of). The underlying conceptual framework is that of the information feedback loop together with resource stocks and flows and an endogenous perspective on causation. The simplicity of the loop concept is apt to contribute to the apparent ease with which SD models can be created (along with the icon-based suites of SD software). But the novice reader should appreciate that it can take time to assimilate the modeling skills necessary to execute well an SD model-based application. Practice is essential and the references included will lead to further published material to assist the steep climb up the learning curve. So-called experts are still being confronted with the subtleties of SD modeling after years of involvement.

To place the SD methodology in context, the contribution by Schwaninger (see ► System Dynamics in the Evolution of the Systems Approach) profiles it alongside various others 'systems' based approaches which have emerged in the management and social sciences. Those professing to become experts in SD need to know about the other range of approaches which co-exist in the field of systems science. All these other methodologies have their own enthusiasts and this may even extend to the formation of societies with annual conferences. His Appendix B shows a diagram of the different systems approaches and their interrelationships.

The foundations of the SD methodology can be characterized by certain philosophical issues. Olaya's text (see ► System Dynamics Philosophical Background and Underpinnings) defines a central one as presentationalism, associated with the notion of 'mental models'. A number of other philosophical issues which relate to SD are introduced, including those of positivism and social theory.

The practice of SD when applied to real-world applications essentially involves managerial learning and will often involve an interaction with client teams rather than one individual. How best to organize such structured approaches to participative model building is described by Rouwette and Vennix (see ► Group Model Building). Client participation is required for successful modeling.

If the promotion of learning and understanding is the primary *raison d'etre* of SD, then achievement of this goal in an individual can be a significant accomplishment, especially if that person is the most senior in the client team. But there is a further goal to be pursued should the study fully reap the benefits of the SD methodology: How can we foster *organizational* learning? Maani tackles this head on (see ► System Dynamics and Organizational Learning). He defines the core capabilities of a learning organization and goes on to list the developing literature on organizational learning and, most importantly, how SD can aid the process through learning laboratories and microworlds.

Running an SD model creates a time-path of output behavior covering all the variables it is deemed necessary to include in the model. The various runs of the model are, most frequently, addressed in comparative fashion rather than taken in isolation. They can therefore be described as computer-based scenarios each of which charts a possible but not assured future. Georgantzas (see ► Scenario-Driven Planning with System Dynamics) describes environmental (traditional) scenario generation for which there is a considerable body of literature. But he emphasizes that successful strategy design involves the integration of three things: a knowledge of the business environment; the effects of unstated assumptions about change in the environment and strategy on performance; and finally the need to *compute* the effects on organizational performance. These three facets are accomplished by the process of SD modeling.

Thus far this introductory roadmap has covered all the background for contextualizing and creating an SD model. We now turn to various tasks associated with ex post modeling activities. Three such aspects are covered: model validation; analytical methods to explain behavior and determine dominant loops; and model optimization.

Schwaninger and Groesser (see ► System Dynamics Modeling: Validation for Quality Assurance) range over the various aspects of model validation, beginning with its epistemological foundations. In real-world modeling studies testing and validation is a sine qua non of the process. The range of tests made available and the attention given to the task of validation in the literature mark out SD as unique in the field of management science. Few other methodologies get near to the variety of tests which can be applied to an SD model. The authors consider the range of tests under three headings: model-related context; model structure; and model behavior.

Kampmann and Oliva deal with the behavioral analysis issue (see ► System Dynamics, Analytical Methods for Structural Dominance Analysis in). This activity tries to shed light on the model's dynamic behavior: Why does it behave as it does? What loop structures are responsible for the dominant behavior – and indeed shifts in that behavior where it occurs? In other words, they explore the link between system structure and dynamic behavior. Early methods used eigenvalue analysis but, since then, more sophisticated approaches have been put forward. A major advance will occur when one or more of these is refined enough to be included in an SD software package. This is likely to take some time although an improved user interface showing links glowing with differing degrees of intensity, reflecting their relative importance, is possible in the not-too-distant future.

Dangerfield describes the methods for improving model performance (see ► System Dynamics Models, Optimization of). The task can be categorized under two headings: calibration and policy optimization. The former relates to the determination of optimal parameter sets which deliver the best fit of the model to past time series data. Policy optimization on the other hand seeks to establish policies which deliver the 'best' performance against a suitable metric, such as minimum cost or maximum revenue. Using such an approach can accelerate the learning which comes from repeated runs of the model. Sadly, in the existing SD literature, there is scant evidence of its use in real-world studies.

The methodology of SD exists for no other reason than to offer a quantum leap in the standards of policy analysis. Therefore, any review must include a range through the landscape which defines areas of application. There are eight such areas covered in this section and the choice has been made in the knowledge that there are others which may also have been included and some new areas which are only just being opened up to the tools of SD modeling.

Business Strategy was the genesis of SD applications and rightly takes pride of place. This is the field in which the most numerous SD applications occur. Lyneis (see ► Business Policy and Strategy, System Dynamics Applications to) concentrates on the process of how SD models are used in the task of strategy formulation. He goes on to consider the various drivers of business dynamics such as oscillations in supply chains and boom and bust life cycles. Detailed references are provided for a wide range of business application case studies.

Health care is consuming a higher share of GDP in many Western industrialized countries. This is due to the age profile of the population and advances in pharmacological and medical technologies. It is unsurprising that SD methods have been applied in tackling some of the most high-profile issues in health care and the relatively recent literature is testimony to the success of SD-based analysis. Indeed, it is arguable that some of the best modeling applications have surfaced in this sector. To do justice to the field of health care two contributions were solicited, in part because of the different funding systems which exist on either side of the Atlantic: Wolstenholme surveys the work done by UK and European authors (see ► Health Care in the United Kingdom and Europe, System Dynamics Applications to), whilst Hirsch and Homer concentrate on work published by US authors (see ► Health Care in the United States, System Dynamics Applications to).

Wolstenholme describes work carried out in the UK and Continental Europe but gives particular emphasis to three areas where models have been deployed. He starts with the problem of delayed hospital discharge which generates hospital capacity problems. Epidemiology is also reviewed, in particular research on the epidemiology of HIV/AIDS. Finally, recent work on mental health reform in the UK is described.

Hirsch and Homer note that the system in the USA is comparatively difficult to manage because of its free-market approach and relative lack of regulation. They concentrate on three main areas: disease epidemiology including heart disease and diabetes; substance abuse; and health care capacity and delivery.

Along with health care, the depletion of environmental resources and its effects has consumed many thousands of column inches in printed news media. SD has been employed in the pursuit of more compelling applications in this sector and the efforts go back to the well-known *Limits to Growth* study in 1971–72. Ford charts the most notable efforts which have emerged (see ► System Dynamics Models of Environment, Energy and Climate Change). He ranges over environmental resource problems in the western USA, models for greater understanding of climate change and global warming and concludes with studies in energy, specifically two applications to the electric power industry.

The field of economics is one where SD has received a mostly hostile reception. The statistical economic modeling tool of econometrics has an extensive history and as a preferred modeling methodology seems hard to dislodge. However, there are an increasing number of heterodox economists who are prepared to embrace SD concepts and Radzicki (see ► System Dynamics and Its Contribution to Economics and Economic Modeling) describes the advances taking place. Whilst some of the literature embodies the translation of existing economic models into an SD format (which is a laudable objective) he calls for more economic dynamics models to be built from scratch embodying the best practice in SD modeling. Economic policy is too important to be informed by a single, seemingly unassailable, modeling methodology and it is to be hoped that in the future SD will become even more accepted as a viable tool for use in this field.

In a similar vein comes the contribution of Saeed (see ► Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation). He takes an economic modeling perspective and describes an SD model which explains resource allocation, production and entitlements in a market economy. Its purpose is to understand better how poverty might be reduced in the context of the redistribution of income. A comprehensive listing of the model is provided in an appendix.

The application of SD to public policy generally is dealt with by Andersen, Rich and MacDonald (see ► Public Policy, System Dynamics Applications to). They emphasize how public policy issues are complex, cross organizational boundaries, involve stakeholders with widely different perspectives and evolve over time, such that longer term results may be wholly different from short-term outcomes. Detail is provided for one public policy case involving the Governor's Office of Regulatory Assistance in New York State. They conclude with coverage of studies in a range of public domains such as defense, health care, education and the environment.

One area of SD application has brought the methodology into the legal arena. Disruption and delay in the execution of complex projects invariably finds two parties in dispute. Such disputes often center upon time delays and use of resources on projects – and what might have happened if things had been managed differently. SD models have been employed by parties to such disputes to attempt to justify the occurrence of these events. Howick, Ackermann, Eden and Williams (see ► Delay and Disruption in Complex Projects) report on how cognitive mapping, cause mapping and SD can be fused into what they describe as a cascade model building process. The result is a rigorous process for explaining why a project behaved in a certain way.

New products and processes are emerging at an ever-increasing rate in modern times. We need to understand the myriad mechanisms which are the basis for their rate of adoption. Milling and Maier range over various SD models which have been created to understand and improve the management of the diffusion of innovations (see ► Diffusion of Innovations, System Dynamics Analysis of the). From the often-cited Bass diffusion model (1969) the authors develop a series of additional features in a modular fashion. These features include competition, network externalities, dynamic pricing and research and development. They conclude by stressing how it is not possible to offer general recommendations for strategies in dynamic and complex environments; such recommendations can only be given in the context of the specific case under scrutiny.

System Dynamics, The Basic Elements of

GEORGE P. RICHARDSON
Rockefeller College of Public Affairs and Policy,
University at Albany, State University of New York,
Albany, USA

Article Outline

Glossary

Endogenous Generated from within. Contrasting with "exogenous," meaning generated by forces external to a system or point of view.

Feedback loop A closed path of causal influences and information, forming a circular-causal loop of information and action.

System dynamics System dynamics is a computer-aided approach to theory-building, policy analysis and strategic decision support emerging from an endogenous point of view.

Definition of the Subject

System dynamics is a computer-aided approach to theory-building, policy analysis, and strategic decision support emerging from an endogenous point of view [18,20]. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems – literally any dynamic systems characterized by interdependence, mutual interaction, information feedback, and circular causality.

Introduction

The field of system dynamics developed initially from the work of Jay W. Forrester. His seminal book *Industrial Dynamics* [7] is still a significant statement of philosophy and methodology in the field. Within ten years of its publication, the span of applications grew from corporate and industrial problems to include the management of research and development, urban stagnation and decay, commodity cycles, and the dynamics of growth in a finite world. It is now applied in economics, public policy, environmental studies, defense, theory-building in social science, and other areas, as well as its home field, management. The name industrial dynamics no longer does justice to the breadth of the field (for extensive examples, see [20,28], so it has become generalized to system dynamics. The modern name suggests links to other systems methodologies, but the links are weak and misleading. System dynamics emerges out of servomechanisms engineering, not general systems theory or cybernetics [18].

The system dynamics approach involves:

- Defining problems dynamically, in terms of graphs over time.
- Striving for an endogenous, behavioral view of the significant dynamics of a system, a focus inward on the characteristics of a system that themselves generate or exacerbate the perceived problem.
- Thinking of all concepts in the real system as continuous quantities interconnected in loops of information feedback and circular causality.
- Identifying independent stocks or accumulations (levels) in the system and their inflows and outflows (rates).
- Formulating a behavioral model capable of reproducing, by itself, the dynamic problem of concern. The model is usually a computer simulation model expressed in nonlinear equations, but is occasionally left unquantified as a diagram capturing the stock-and-flow/causal feedback structure of the system.
- Deriving understandings and applicable policy insights from the resulting model.
- Implementing changes resulting from model-based understandings and insights.

Mathematically, the basic structure of a formal system dynamics computer simulation model is a system of coupled, nonlinear, first-order differential (or integral) equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \mathbf{p}) ,$$

where $\mathbf{x}$ is a vector of levels (stocks or state variables), $\mathbf{p}$ is a set of parameters, and $\mathbf{f}$ is a nonlinear vector-valued function. Such a system has been variously called a *state-determined system* in the engineering literature, an *absolute system* [3], an *equifinal system* [32], and a *dynamical system* [16].

Simulation of such systems is easily accomplished by partitioning simulated time into discrete intervals of length $\mathrm{d}t$ and stepping the system through time one $\mathrm{d}t$

at a time. Each state variable is computed from its previous value and its net rate of change $x'(t)$: $x(t) = x(t - \mathrm{d}t) + \mathrm{d}t \cdot x'(t - \mathrm{d}t)$. In the earliest simulation language in the field (DYNAMO) this equation was written with time scripts K (the current moment), J (the previous moment), and JK (the interval between time J and K): $X_\mathrm{K} = X_\mathrm{J} + \mathrm{DT} \cdot \mathrm{XRATE}_\mathrm{JK}$ (see, e. g., [22]). The computation interval dt is selected small enough to have no discernible effect on the patterns of dynamic behavior exhibited by the model. In more recent simulation environments, more sophisticated integration schemes are available (although the equation written by the user may look like this simple Euler integration scheme), and time scripts may not be in evidence. Important current simulation environments include STELLA and iThink (isee Systems, http://www.iseesystems.com/), Vensim (Ventana Systems, http://www.vensim.com/), and Powersim (http://www.powersim.com/).

Forrester's original work stressed a continuous approach, but increasingly modern applications of system dynamics contain a mix of discrete difference equations and continuous differential or integral equations. Some practitioners associated with the field of system dynamics work on the mathematics of such structures, including the theory and mechanics of computer simulation, analysis and simplification of dynamic systems, policy optimization, dynamical systems theory, and complex nonlinear dynamics and deterministic chaos.

The main applied work in the field, however, focuses on understanding the dynamics of complex systems for the purpose of policy analysis and design. The conceptual tools and concepts of the field – including feedback thinking, stocks and flows, the concept of feedback loop dominance, and an endogenous point of view – are as important to the field as its simulation methods.

Feedback Thinking

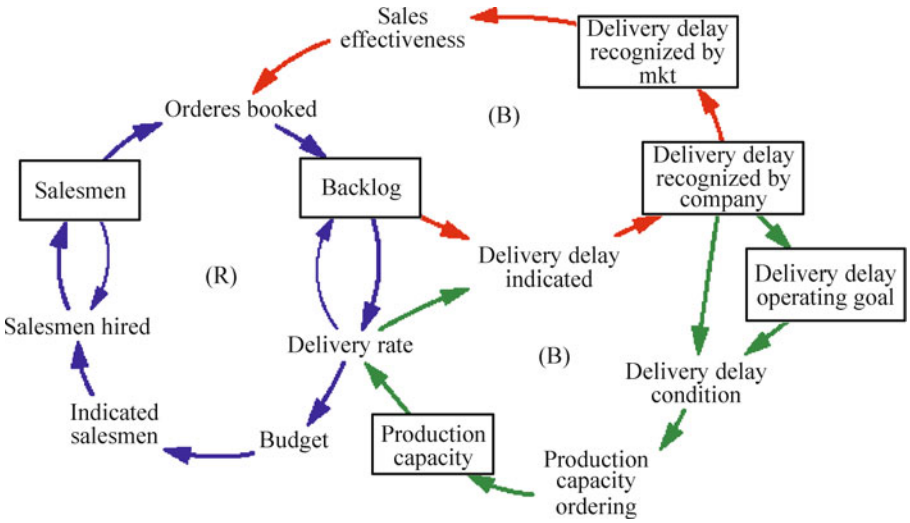
Conceptually, the feedback concept is at the heart of the system dynamics approach. Diagrams of loops of information feedback and circular causality are tools for conceptualizing the structure of a complex system and for communicating model-based insights. Intuitively, a feedback loops exists when information resulting from some action travels through a system and eventually returns in some form to its point of origin, potentially influencing future action. If the tendency in the loop is to reinforce the initial action, the loop is called a *positive* or *reinforcing* feedback loop; if the tendency is to oppose the initial action, the loop is called a *negative, counteracting*, or *balancing* feedback loop. The sign of the loop is called its *polarity*. Balancing loops can be variously characterized as goal-seeking, equilibrating, or stabilizing processes. They can sometimes generate oscillations, as when a pendulum seeking its equilibrium goal gathers momentum and overshoots it. Reinforcing loops are sources of growth or accelerating collapse; they are disequilibrating and destabilizing. Combined, balancing and reinforcing circular causal feedback loops can generate all manner of dynamic patterns.

Feedback loops are ubiquitous in human and natural systems and, under various names and representations, have been widely recognized in popular and scholarly literature. Feedback thought has been present implicitly or explicitly for hundreds of years in the social sciences and literally thousands of years in recorded history [9]. We have the vicious circle originating in classical logic and morphing into common usage, the bandwagon effect, the invisible hand of Adam Smith, Malthus's correct observation of population growth as a self-reinforcing process, Keynes's consumption multiplier, the investment accelerator of Hicks and Samuelson, compound interest or inflation, the biological concepts of proprioception and homeostasis, Festinger's cognitive dissonance, Myrdal's principle of cumulative causation, Venn's idea of a suicidal prophecy, Merton's related notion of a self-fulfilling prophecy, and so on. Each of these ideas can be concisely and insightfully represented as one or more loops of causal influences with positive or negative polarities. Great social scientists and feedback thinkers; great social theories are feedback thoughts. (For a full exposition of the evolution of the feedback concept see [19].)

Loop Dominance and Nonlinearity

The loop concept underlying feedback and circular causality by itself is not enough, however. The explanatory power and insightfulness of feedback understandings also rest on the notions of active structure and loop dominance. Complex systems change over time. A crucial requirement for a powerful view of a dynamic system is the ability of a mental or formal model to change the strengths of influences as conditions change, that is to say, the ability to shift *active* or *dominant structure*.

In a system of equations, this ability to shift loop dominance comes about endogenously from nonlinearities in the system. For example, the S-shaped dynamic behavior of the classic logistic growth model ($\mathrm{d}P/\mathrm{d}t = aP - bP^2$) or similar structures like the Gompertz curve ($\mathrm{d}P/\mathrm{d}t = aP - bP \ln(P)$) can be seen as the consequence of a shift in loop dominance from a positive, self-reinforcing feedback loop (aP) producing exponential-like growth, to a negative



System Dynamics, The Basic Elements of, Figure 1
Core structure of Forrester's market growth model [8], showing a *blue reinforcing loop* underlying the growth (or reinforcing decline) of Salesmen, Orders, and Revenue, a *red balancing loop* containing various delayed recognitions of the company's delivery delay, and a *green balancing loop* responsible for capacity ordering if the delivery delay drops too far below its operating goal

feedback loop ($-bP^2$ or $-bP\ln(P)$) that brings the system to its eventual goal. The shift in loop dominance in these models comes about from the nonlinearity in the second term, which grows faster than the first term and eventually overtakes it. Only nonlinear models can endogenously alter their active or dominant structure and shift loop dominance.

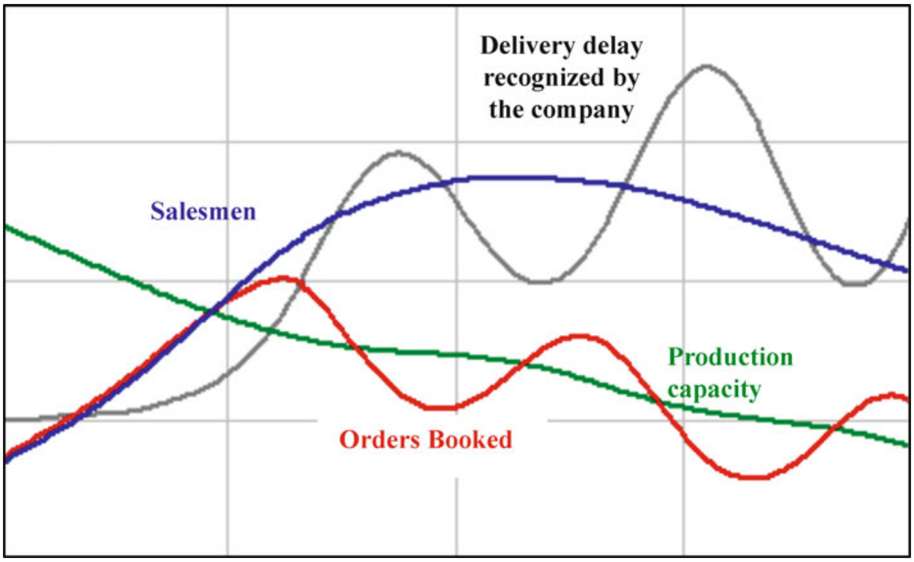
Real systems are perceived to change their active or dominant structure over time, often because of the build-up of internal forces. Thus from a feedback perspective, the ability of nonlinearities to generate shifts in loop dominance is the fundamental reason for advocating nonlinear models of social system behavior.

Figures 1 and 2, abstracted from an early, classic paper [8] illustrate these ideas. In Fig. 1 salesmen (in the blue reinforcing loop) book orders for the company; if enough revenue is generated, there is enough budget to hire more salesmen and corporate growth ensues. Whether salesmen (in this simplified picture) book enough orders depends on the company's delivery delay for the product, as perceived by the market (red balancing loop). The company builds production capacity according to its perceived need, as indicated by its perceived delivery delay and its target for that (green balancing loop).

Figure 2 shows the dynamics this feedback structure endogenously generates. In the early phase, salesmen grow as orders and revenue grow; the system's exponential growth behavior in that phase is generated by the reinforcing salesmen loop. But then the feedback loop dominance soon shifts to the balancing delivery delay loop, which constrains sales effectiveness and brings a halt to growth. The system moves into an oscillatory phase generated by the various monitoring and perception delays around the now dominant red balancing loop. Salesmen eventual peak and decline, as the green production capacity ordering loop fails to keep production capacity sufficient to hold the delivery delays in check.

Thus the dynamic behavior of this system is a consequence of its feedback structure and the nonlinearities that shift loop dominance endogenously over time. The particular decline scenario shown in Fig. 2 illustrates one of the deep insights of the model: the adaptive goal structure, in which the delivery delay operating goal moves slowly to accommodate changes in the company's delivery delay, weakens the green balancing loop trying to bring on capacity. The company never perceives its delivery delay is sufficiently higher than its (sliding) target, so it fails to order sufficient capacity to sustain growth. A fixed goal for the acceptable delivery delay sends a stronger signal, which can turn this corporate decline into oscillating growth [8].

Thus, nonlinearity is crucial to the system dynamics approach. However, it is crucial not merely because of its mathematical properties but because it enables the formalization of a profoundly powerful perspective on theory and policy – the *endogenous point of view*.



System Dynamics, The Basic Elements of, Figure 2
The dynamic behavior of the model shown in Fig. 1, illustrating an early growth phase, which turns into an oscillatory phase as the feedback loop dominance shifts to the *red balancing delivery delay loop*, and results in a long term corporate decline as the *green capacity ordering loop* responds to a sliding operating goal for the acceptable delivery delay

The Endogenous Point of View

The concept of endogenous change is fundamental to the system dynamics approach. It has both philosophical and engineering origins. A deep and lasting insight of the earliest attempts at servomechanisms control is the realization that *the attempt to control a system generates dynamics of its own*, complicating the dynamics trying to be controlled. A governor mechanism imposed to control the speed of a steam engine can generate oscillatory "hunting behavior," as the control system overshoots and undershoots the set point. As it becomes part of the system, the governing mechanism thus generates dynamics of its own.

The insight transfers readily, but with added significance, from engineering systems to people systems: Attempts to control complex human systems – coercing, guiding, managing, governing – generate dynamics of their own. Moreover, some of these endogenously generated dynamics are created by the control mechanisms themselves (like the governor of a steam engine) and some are created by human creative responses to the management efforts (e. g., principal-agent interactions). These natural and human forces, creating counteracting and compensating pressures in response to system control efforts, emerge as complicated circular-causal feedback structures. The often complex, difficult-to-understand dynamics of such management systems are to a great degree a consequence of their internal structures.

To capture and analyze such management complexities, one must look inward to see the ways a complex system naturally responds to system pressures. The endogenous point of view is thus central to the system dynamics approach. It dictates aspects of model formulation: exogenous disturbances are seen at most as *triggers* of system behavior (like displacing a pendulum); the *causes* are contained within the structure of the system itself (like the interaction of a pendulum's position and momentum that produces oscillations). Corrective responses are also not modeled as functions of time, but are dependent on conditions within the system. Time by itself is not seen as a cause in the endogenous point of view.

Theory building and policy analysis are significantly affected by this endogenous perspective. Taking an endogenous view exposes the natural *compensating* tendencies in social systems that conspire to defeat many policy initiatives. Feedback and circular causality are delayed, devious, and deceptive. For understanding, system dynamics practitioners strive for an *endogenous point of view*. The effort is to uncover the sources of system behavior that exist within the structure of the system itself.

System Structure

These ideas are captured almost explicitly in Forrester's [9] organizing framework for system structure:

- Closed boundary
- Feedback loops
- Levels
- Rates
- Goal
- Observed condition
- Discrepancy
- Desired action.

The *closed boundary* signals the endogenous point of view. The word *closed* here does not refer to open and closed systems in the general system sense, but rather refers to the effort to view a system as *causally* closed. The modeler's goal is to assemble a formal structure that can, *by itself*, without exogenous explanations, reproduce the essential characteristics of a dynamic problem.

The causally closed system boundary at the head of this organizing framework identifies the endogenous point of view as the feedback view pressed to an extreme. Feedback thinking can be seen as a *consequence* of the effort to capture dynamics within a closed causal boundary. Without causal loops, all variables must trace the sources of their variation ultimately outside a system. Assuming instead that the causes of all significant behavior in the system are contained within some closed causal boundary forces causal influences to feed back upon themselves, forming causal loops. Feedback loops enable the endogenous point of view and give it structure.

Levels and Rates

Stocks (accumulations, or "levels" in early system dynamics literature) and the flows ("rates") that affect them are essential components of system structure. A map of causal influences and feedback loops is not enough to determine the dynamic behavior of a system. A constant inflow yields a linearly rising stock; a linearly rising inflow yields a stock rising along a parabolic path; a stock with inflow proportional to itself grows exponentially; two stocks in a balancing loop have a tendency to generate oscillations; and so on. For example, the boxes in Fig. 1 represent accumulations in the company and its market; the three stocks in the red balancing loop (the order backlog and the two perceptions of the company's delivery delay) give that loop its tendency to generate oscillations which propagate through out the system. Accumulations are the memory of a dynamic system and contribute to its disequilibrium and dynamic behavior.

Forrester [7] placed the operating policies of a system among its rates, the inflows and outflows governing change in the system. Many of these rates of change assume the classic structure of a negative feedback loop striving to take action to reduce the discrepancy between the observed condition of the system and a goal. The simplest such rate structure results in an equation of the form

$$\text{RATE} = \frac{\text{GOAL} - \text{LEVEL}}{\text{ADJUSTMENT TIME}} ,$$

where ADJUSTMENT TIME is the time over which the level adjusts to reach the goal. This simple formulation reflects Forrester's more general statement about rates in his hierarchy of system structure (above) which can be richly thought of as

$$\text{RATE} = f(\text{DESIRED ACTION})$$

$$\begin{aligned}&\text{DESIRED ACTION}\\&\quad = g(\text{DESIRED CONDITION},\\&\qquad\qquad\qquad\qquad \text{OBSERVED CONDITION})\end{aligned}$$

$$\text{OBSERVED CONDITION} = h(\text{LEVELS}) ,$$

for some functions f, g, and h representing particular system characteristics.

Operating policies in a management system can influence the *flows* of information, material, and resources, which are the only means of changing the accumulations in the system. While flows can be changed quickly, as a matter of relatively quick decision making, stocks change slowly – they rise when inflows are great than outflows, and decline when inflows are less than outflows.

The simple "tub dynamics" of stocks are clear even to children, yet can be befuddling in complex systems. The accumulation of green house gases in the atmosphere, for example, affects the *flow* of heat energy radiated from the earth. To turn around global warming, the *accumulation* of green house gases must drop far enough to raise radiant energy above the inflow of solar energy, a simple stock-and-flow insight. But to cause the accumulation of green house gases to drop, their generation must fall below their natural absorption rate (another simple stock-and-flow observation). So turning around global warming is a process involving a chain of at least two significant accumulations, and people have trouble thinking it through reliably. The accumulations can only be changed by managing their associated flows. They will change only slowly even if we manage the technical and political pitfalls involved in lowering green house gas production (see [29]).

The significance of stocks in complex systems is vivid in a resource-based view of strategy and policy. Resources that enable a corporation or government to function or flourish are stocks, usually accumulated over long periods of time with significant investment of time, energy, and money. Reputations are also stocks, built over similarly long periods of time. While inadequate by themselves to give a full picture of the dynamics of a complex system, stocks and flows are vital components of system structure, without which fundamental understandings of dynamics are impossible [33].

Behavior is a Consequence of System Structure

The importance of stocks and flows appears most clearly when one takes a *continuous* view of structure and dynamics. Although a discrete view, focusing on separate events and decisions, is entirely compatible with an endogenous feedback perspective, the system dynamics approach emphasizes a continuous view [7]. The continuous view strives to look beyond events to see the dynamic patterns underlying them: model not the appearance of a discrete new housing unit in a city, but focus instead on the rise and fall of aggregate numbers of housing units. Moreover, the continuous view focuses not on discrete decisions but on the *policy structure* underlying decisions: not why this particular apartment building was constructed but what persistent pressures exist in the urban system that produce decisions that change housing availability in the city. Events and decisions are seen as surface phenomena that ride on an underlying tide of system structure and behavior. It is that underlying tide of policy structure and continuous behavior that is the system dynamicist's focus.

There is thus a *distancing* inherent in the system dynamics approach – not so close as to be confused by discrete decisions and myriad operational details, but not so far away as to miss the critical elements of policy structure and behavior. Events are deliberately blurred into dynamic behavior. Decisions are deliberately blurred into perceived policy structures. Insights into the connections between system structure and dynamic behavior, which are the goal of the system dynamics approach, come from this particular distance of perspective.

Suggestions for Further Reading on the Core of System Dynamics

The *System Dynamics Review*, the journal of the System Dynamics Society, published by Wiley, is the best source of current activity in the field, including methodological advances and applications.

The core of a vibrant field is difficult to discern in the flow of current work. However, the works that the field itself singles out as exemplary can give some reliable hints about what is considered vital to the core. In this sense two edited volumes are noteworthy: An early, interesting collection of applications is Roberts [24]; Richardson [21] is a more recent two-volume edited collection in the same spirit, containing prize-winning work in philosophical background, dynamic decision making, applications in the private and public sectors, and techniques for modeling with management.

In addition, the following works, selected from among winners of the System Dynamics Society's *Jay Wright Forrester Award* (see www.systemdynamics.org/Society_Awards.htm), can be considered insightful although implicit exemplars of the core of system dynamics. (Publications are listed beginning with the most recent; see the bibliography for full citations):

- Thomas S. Fiddaman, "Exploring policy options with a behavioral climate-economy model"
- Kim D. Warren, *Competitive Strategy Dynamics*
- Eric F. Wolstenholme, "Towards the Definition and Use of a Core Set of Archetypal Structures in System Dynamics"
- Nelson P. Repenning, "Understanding Fire Fighting in New Product Development"
- John D. Sterman, *Business Dynamics, Systems Thinking and Modeling for a Complex World*
- Peter Milling, "Modeling innovation processes for decision support and management simulation."
- Erling Moxnes, "Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics."
- Jac A. M. Vennix, *Group Model Building: Facilitating Team Learning Using System Dynamics*
- Jack B. Homer, "A System Dynamics Model of National Cocaine Prevalence."
- Andrew Ford, "Estimating the Impact of Efficiency Standards on Uncertainty of the Northwest Electric System."
- Khalid Saeed, *Towards Sustainable Development: Essays on System Analysis of National Policy*
- Tarek Abdul-Hamid and Stuart Madnick, *Software Project Dynamics: An Integrated Approach*
- George P. Richardson, *Feedback Thought in Social Science and Systems Theory*
- Peter M. Senge, *The Fifth Discipline*
- John D. W. Morecroft, "Rationality in the Analysis of Behavioral Simulation Models."
- John D. Sterman, "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment."

For texts on the system dynamics approach, see Alfeld and Graham [2], Richardson and Pugh [22], Wolstenholme

[34], Ford [6], Maani and Cavana [11], and the most comprehensive text to date, Sterman [28].

Bibliography

1. Abdul-Hamid T, Madnick S (1991) Software project dynamics: an integrated approach. Prentice Hall, Englewood Cliffs
2. Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Pegasus Comunications, Waltham
3. Ashby H (1956) An Introduction to cybernetics. Chapman & Hall, London
4. Fiddaman TS (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
5. Ford A (1990) Estimating the impact of efficiency standards on uncertainty of the northwest electric system. Oper Res 38(4):580–597
6. Ford A (1999) Modeling the environment: an introduction to system dynamics of environmental systems. Island Press, Washington
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
8. Forrester JW (1968) Market growth as influenced by capital investment. Ind Manag Rev (MIT, now Sloan Manag Rev) 9(2):83–105. Reprinted widely, e. g., Richardson [21]
9. Forrester JW (1969) Urban dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
10. Homer JB (1992) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78
11. Maani KE, Cavana RY (2000) Systems thinking and modelling: understanding change and complexity. Pearson Education, New Zealand
12. Milling P (1996) Modeling innovation processes for decision support and management simulation. Syst Dyn Rev 12(3):211–234
13. John DW, Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31(7):900–916
14. Morecroft JDW, Sterman JD (eds) (1994) Modeling for learning organizations. System dynamics series. Pegasus Communications, Waltham
15. Moxnes E (1998) Not only the tragedy of the commons: misperceptions of bioeconomics. Manag Sci 44(9):1234–1248
16. Nicholis G, Prigogine I (1977) Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations. Wiley, New York
17. Repenning NR (2001) Understanding fire fighting in new product development. J Prod Innov Manag 18(5):285–300
18. Richardson GP (1991) System dynamics: simulation for policy analysis from a feedback perspective. In: Fishwick PA, Luker PA (eds) Qualitative simulation modeling and analysis. Springer, New York
19. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia. Reprinted by Pegasus Communications, 1999
20. Richardson GP (1996) System dynamics. In: Gass S, Harris C (eds) The encyclopedia of operations research and management science. Kluwer, New York
21. Richardson GP (ed) (1996) Modelling for management: simulation in support of systems thinking. International library of management. Dartmouth, Aldershot
22. Richardson GP, Pugh AL III. (1981) Introduction to system dynamics modeling with DYNAMO. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
23. Richmond B (1993) Systems thinking: critical thinking skills for the 1990s and beyond. Syst Dyn Rev 9:(2)113–133
24. Roberts EB (ed) (1978) Managerial applications of system dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
25. Saeed K (1991) Towards sustainable development: essays on system analysis of national policy. Progressive Publishers, Lahore
26. Senge PM (1990) The fifth discipline: the art and practice of the learning organization. Doubleday/Currency, New York
27. Sterman JD (1988) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339
28. Sterman JD (2001) Business dynamics: systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
29. Sterman JD, Sweeney LB (2002) Cloudy skies: assessing public understanding of global warming. Syst Dyn Rev 18(2):207–240
30. System Dynamics Review. 1985–present. Wiley, Chichester
31. Vennix JAM (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
32. von Bertalanffy L (1968) General systems theory: foundations, development, applications. George Braziller, New York
33. Warren KD (2002) Competitive strategy dynamics. Wiley, United Kingdom
34. Wolstenholme EF (1990) System enquiry: a system dynamics approach. Wiley, Chichester
35. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

Treasury Market, Microstructure of the U.S.

BRUCE MIZRACH[1], CHRISTOPHER J. NEELY[2]
[1] Department of Economics, Rutgers University, New Brunswick, USA
[2] Research Department, Federal Reserve Bank of St. Louis, St. Louis, USA

Article Outline

Glossary

Algorithmic trading Algorithmic trading is the practice of automatically transacting based on a quantitative model.

Broker A broker is a firm that matches buyers and sellers in financial transactions. An *interdealer broker (IDB)* is an intermediary providing trading services to hedge funds, institutions, and other dealers. IDB's handle the majority of Treasury securities transactions in the secondary market.

Coupons Owners of Treasury notes and bonds receive periodic payments called coupons. They are fixed by the Treasury at auction and are typically paid semi-annually.

Depth Depth is the quantity the dealer is willing to sell at the bid or offer.

Electronic communications networks (ECN) The Securities and Exchange Commission defines electronic communications networks (ECNs) as "electronic trading systems that automatically match buy and sell orders at specified prices".

Market microstructure Market microstructure is a field of economics that studies the price formation process and trading procedures in security markets.

On-the-run On-the-run refers to the most recently auctioned Treasury security of a particular maturity. After the next auction, the security goes *off-the-run*.

Price discovery The process by which prices adapt to new information.

Primary dealers Primary dealers are large brokerage firms and investment banks that are permitted to trade directly with the Federal Reserve in exchange for making markets in Treasuries. They provide the majority of liquidity in the Treasury market, participate in Treasury auctions, and provide information to assist the Fed in implementing open market operations.

Secondary market After the initial auction of Treasury instruments, trading in on-the-run and off-the-run securities makes up the *secondary* Treasury market.

When issued When-issued bonds are those Treasuries whose auctions have been announced but have not yet settled.

Definition of the Subject

This article discusses the microstructure of the *US Treasury securities market*.

US Treasury securities are default risk free debt instruments issued by the US government. These securities play an important, even unique, role in international financial markets because of their safety, liquidity, and low transactions costs. Treasury instruments are often the preferred safe haven during financial crises, a process often referred to as a "flight to quality".

According to the US Treasury, there was more than $9 trillion in US government debt outstanding as of August 31, 2007. Of this quantity, the public holds more than $5 trillion and $4.5 trillion is tradable on financial markets. Foreigners hold approximately $2.4 trillion of the marketable supply, with Japan and China together holding more than $1 trillion. According to the Securities Industry and Financial Markets Association (SIFMA), average daily trading volume in the US Treasury market in 2007 was $524.7 billion.

Microstructure is the study of the institutional details of markets and trading behavior. Microstructural analysis takes three ideas seriously that are often overlooked: the institutional features of the trading process influence how private information is impounded into prices; agents are heterogeneous; and information is asymmetric. Empirical microstructure research studies topics such as the causes and effects of market structure, how market structure influences price discovery, how trading and order flow reveal private information, how quickly public information is impounded into prices, the volatility-volume relation, and the determinants of transactions costs (i. e., the components of bid-ask spreads). The relatively recent availability of tick-by-tick financial data and limit order book data, as well as the computer resources to manipulate them, have been a great boon to financial market microstructure research.

Introduction

We begin by describing the types of Treasury issues and the major Treasury market participants, including the Federal Reserve, primary dealers and the major electronic brokers. We then outline the stages of the Treasury market, from auction announcements to the secondary market. Next, we examine several closely related areas of the literature: Seasonality in the Treasury market and the reactions of the Treasury market to macro and monetary announcements; discontinuities in Treasury prices; and the effect of order flow in Treasury markets. We then discuss modeling and other academic questions about the Treasury market.

Types of Treasury Issues

As of October 2007, the US Treasury issued four types of debt instruments. The shortest-maturity instruments are known as Treasury *bills*. 22.6% of the marketable US debt is in bills, securities with maturities of 1 year or less. Bills are sold at a discount and redeemed at their face value at maturity. They do not pay any coupons prior to maturity and currently have maturities up to 26 weeks. Treasury bill prices are usually quoted in "discount rate" terms, which are calculated with an actual/360 day count convention,

$$\text{T-bill discount rate} = [\text{face value} - \text{bill price}] \times (360/\text{number of days until maturity})\,.$$

Thus, a bill with a face value of $100,000, a cash price of $97,500 and 90 days to maturity will have a discount rate of 10% $= [100-97.5]\times(360/90)$ in a newspaper. Treasury bill yields are often quoted as "bond equivalent yields", which are defined as,

$$\text{T-bill yield} = \left[\frac{\text{face value} - \text{bill price}}{\text{bill price}}\right] \times (365/\text{number of days until maturity})\,.$$

Treasury instruments with intermediate maturities (2-, 5- and 10-year) are known as *Treasury notes*. *Notes* pay semi-annual coupons, and make up 54.7% of the debt. In February 2006, the US Treasury also resumed issuing 30-year instruments, known as *Treasury bonds*. *Bonds* also pay semi-annual coupons, and make up 12.5% of the US debt.

The price of both notes and bonds are quoted as a percentage of their face value in thirty-seconds of a point. A quoted price of 98-08 means that the quoted price of the note (or bond) is $(98 + 8/32 \;=)$ $98.25 for each $100 of face value. The cash price of bonds and notes is equal to the quoted price plus accrued interest since the last coupon payment, calculated with an actual/actual day count convention. Quoted prices are sometimes called "clean" prices, while cash prices are said to be "dirty".

The US Treasury also issues 5-, 10-, and 20-year Treasury Inflation–Protected Securities ("TIPS"), whose payoff is linked to changes in the US Consumer Price Index (CPI). These make up about 10.2% of the total value of Treasuries outstanding. The principal value of TIPS is adjusted daily and the semi-annual coupon payments and principal payment are then based on the adjusted principal amount. Economists extract inflation forecasts by comparing the TIPS yields to those on similar nominal instruments. The Federal Reserve Bank of Saint Louis provides "TIPS spreads" through its publication, *Monetary Trends*.

There is also an active market in STRIPS (Separate Trading of Registered Interest and Principal of Securities) which are popularly known as "zero coupon" bonds. These instruments are created by the Treasury through an accounting system which separates coupon interest payments and principal. Finally, the US Treasury also issues savings bonds, low denomination securities for retail investors.

Treasury Market Participants

The Federal Reserve in the Treasury Market

The Federal Reserve Bank of New York, under the guidance of the Federal Open Market Committee (FOMC), is a uniquely important player in the Treasury market. The FOMC meets approximately every six weeks to review economic conditions and determine a target for the federal funds rate, the rate at which US banks borrow/lend reserve balances from/to each other. The manager of the Open Market Desk (a.k.a., "the Desk") at the Federal Reserve Bank of New York is responsible for ensuring that the average federal funds transaction is close to the target by buying and selling Treasury instruments (primarily short-term). In practice, the Desk accomplishes this in two ways. First the Desk buys sufficient Treasuries to satisfy most but not all the markets' demand for deposits at the Fed. Secondly, the Desk buys Treasuries via repurchase (repos) agreements (overnight and for terms of several days) to achieve a desired repo rate that influences the federal funds rate and other short-term interest rates through arbitrage.

To determine day-to-day actions, every morning, staff at both the Division of Monetary Affairs of the Board of Governors of the Federal Reserve System and the Desk forecast that day's demand for reserve balances. The Desk staff also consults market participants to get their views on financial conditions. The relevant Desk and Board staffs then exchange views in a 9 am conference call. Finally, the relevant Desk staff, the Board staff, and at least one of the voting Reserve Bank Presidents then confer during a second conference call at about 9:20 am. The Desk staff summarizes market conditions, projects actions for the day and asks the voting Reserve Bank President(s) for comments. Open market operations commence shortly after the conclusion of this call.

When the Desk buys Treasuries, it increases available liquidity (reserves) in debt markets and tends to lower interest rates. Selling Treasuries has the opposite effect, lowering reserves and raising interest rates. If the intention is to make a permanent change in reserves, then outright purchases or sales are undertaken. In contrast, if the Desk anticipates that only temporary changes in reserves are necessary, it uses repos (for purchases) or reverse repos (for sales). Bernanke [7] notes that actual open market sales of debt instruments are rare; it is more common for the Federal Reserve to allow such securities to expire without replacing them. Both open market sales and allowing the Fed's securities to expire have the same balance sheet effects: The Fed holds fewer bonds and more cash, while the public will hold more bonds and less cash.

The Federal Reserve provides several valuable references on its operating procedures. The Annual Report of the Markets Group of the Federal Reserve Bank of New York describes open market operations and current procedures (Federal Reserve Bank of New York, Markets Group [26]). Meulendyke [57] provides a comprehensive view of Federal Reserve monetary policy operations with a historical perspective. Akhtar [1] explains how monetary policy is decided and how such policies affect the economy. Finally, Harvey and Huang [43] gives some historical perspective on operating procedures in the 1980s.

Primary Dealers

Among the most important private sector players in the Treasury markets are the 21 *primary dealers*. The Federal Reserve Bank of New York explains that primary dealers must "participate meaningfully in both the Fed's open market operations and Treasury auctions and … provide the Fed's trading desk with market information and analysis that are helpful in the formulation and implementation of monetary policy". The Federal Reserve does not regulate primary dealers, but does subject them to capital requirements. The Federal Reserve can withdraw a firm's primary dealer designation if it fails to participate in auctions or open market operations or if its capital reserves fall below desired levels.

Interdealer Brokers

Prior to 2000, voice-assisted brokers dominated secondary market trading in Treasuries. Except for Cantor–Fitzgerald, all these brokers reported their trading activity to GovPX, a consortium. In the face of demands by the Securities and Exchange Commission and bond market dealers for greater transparency, five IDBs formed GovPX as a joint venture in 1991. In March 1999, Cantor–Fitzgerald opened up its internal electronic trading platform, eSpeed, to clients. The eSpeed system quickly grabbed a dominant market share, and Cantor Fitzgerald spun off eSpeed as a public company in December 1999. In 2000, a competing electronic brokerage, BrokerTec, joined the market. As in foreign exchange and equity markets, most interdealer and institutional trading in Treasuries quickly migrated from voice networks to these electronic communications networks (ECNs), which have dominated trading in Treasury instruments since 2001. Mizrach and Neely [58] describe the transition from voice assisted trading, largely through the primary dealers, to electronic trading in the Treasury market.

As of November 2007, the two dominant ECNs are eSpeed and BrokerTec. London-based ICAP, PLC, owns BrokerTec while eSpeed merged in the summer of 2007 with BGC, another London based interdealer brokerage. eSpeed and ICAP compete for both on- and off-the-run liquidity. Hilliard Farber and Tullett–Prebon hold the largest brokerage share outside of the dominant two platforms.

Stages of the Treasury Bond Market

The sale of Treasuries undergoes four distinct phases: when issued, primary, on-the-run and off-the-run. Each of these stages has a distinct market structure.

The Primary Market

In the *primary* market, the US Treasury sells debt to the public via auction. The US Treasury usually publishes a calendar of upcoming tentative auction dates on the first Wednesday of February, May, August, and November and bids may be submitted up to 30 days in advance of the auction. In practice, however, the Treasury only announces firm auction information several days in advance and most bids are submitted at that time. Since August 8, 2002, the Treasury has made auction announcements (for all new securities) at 11:00 am Eastern Time (ET). 13- and 26-week bills are auctioned weekly; 2- and 5-year notes are auctioned monthly; 10-year notes are auctioned eight times a year. 30-year bonds, which were reintroduced on February 9, 2006 after a five year hiatus, are auctioned four times a year.

The US Treasury has used a single price auction exclusively since November 1998. Garbade and Ingber [35] discuss the transition from multiple price auctions to the current format single price auctions. All securities are allocated to bidders at the price that, in the aggregate, will result in the sale of the entire issue. This mitigates the risk of a "buyer's curse" – the highest bidder paying more than other auction participants. To prevent a single large buyer from manipulating the auction, the Treasury restricts anyone from buying more than 35% of any single issue. Bids may be submitted up to thirty days prior to the auction, and large institutions make use of the Treasury Automated Auction Processing System (TAAPS). Retail investors can participate through the Treasury Direct program. The Treasury allocates a portion of nearly every auction to small investors at the same price as the large institutions. These are called *non-competitive bids*, and they are quantity only orders that are filled at the market clearing price.

Primary dealers dominate the auction process. In 2003, they submitted 86% of auction bids, totalling more than $6 trillion. They were awarded $2.4 trillion, or 78% of the total auction supply.

The Secondary Market

The secondary market is composed of the when-issued, on-the-run and off-the-run issues.

When-Issued Even prior to the primary auction, there is an active forward market in Treasury securities (apart from TIPS) that are about to be issued. Trading in the *when-issued* security market typically begins several days prior to an auction and continues until settlement of auction purchases. Nyborg and Sundaresan [61] document that when-issued trading provides important information about auction prices prior to the auction and also permits market participants to reduce the risk they take in bidding. Fabozzi and Fleming [25] estimate that 6% of total interdealer trading is in the when-issued market. Just prior to auctions though, these markets become substantially more active. In the bill market, when-issued trading volume exceeds the volume for the bills from the previous auction.

On-the-Run Upon completion of the auction, the most recently issued bill, note or bond becomes *on-the-run* and the previous on-the-run issue goes *off-the-run*. Overall Treasury trading volume is concentrated in a small number of on-the-run issues. Trading in these benchmark on-the-run issues, which Fabozzi and Fleming [25] say constitutes approximately 70% of total trading volume, has migrated almost completely to the electronic networks. Mizrach and Neely [58] estimate a 61% market share for the BrokerTec platform and a 39% share for eSpeed in 2005, which is consistent with industry estimates.

Off-the-Run With more than 200 off-the-run issues trading in October 2007 – 44 bills, 116 notes, and 45 bonds – most off-the-run volume takes place in voice and electronic interdealer networks. Barclay, Hendershott and Kotz [5] document the fall in ECN market share when issues go off the run. They also report that transaction volume falls by more than 90%, on average, once a bond goes off-the-run. The ECN market share falls from 75.2% to 9.9% for the 2-year notes, from 83.5% to 8.5% for the 5-year notes, and from 84.5% to 8.9% for the 10-year notes. Several IDBs handle most off-the-run securities trading.

On- Versus Off-the-Run Liquidity and Prices Off-the-run securities trade at a higher yield (lower price) than on-the-run securities of similar maturity. Many researchers have attempted to explain the yield differential with relative liquidity. Vayanos and Weill [68] utilize a search theoretic model that is motivated by the fact that bonds may be difficult to locate once they go off-the-run. Goldreich, Hanke, and Nath [36] compare on-the-run and off-the-run Treasuries and show that the liquidity premium depends primarily on the amount of remaining future liquidity, which is highly predictable. The study exploits the fact that the liquidity of a Treasury is predictable. Duffie [18] argues that legal or institutional restrictions on supplying collateral induces "special" repo rates that are much less than market riskless interest rates. The price of the underlying instrument is increased by the present value of the savings in borrowing costs.

Supply Variation and Prices Although it is generally accepted that the on-the-run premium is due to greater liquidity, the theoretical relation between the supply of a given bond issue and prices is not clear. Do issue sizes produce lower yields (higher prices) through their liquidity effects or does downward-sloping demand for individual securities produce higher prices (lower yields) for larger issues? Empirically, the evidence is mixed. Simon [65,66], Duffie [18], Seligman [64] and Fleming [29] find that the larger issues lead to lower prices (higher yields), while Amihud and Mendelson [2], Kamara [51], Warga [69], and Elton and Green [23] find the opposite: The liquidity effect predominates, resulting in higher prices (lower yields) for larger issues. There might be a nonlinear relationship. Liquidity may increase prices up to a certain point, but then finite demand for any individual security reduces the attractiveness of additional supply.

The Treasury Futures Market

Spot markets are not the only markets for US Treasuries. The Chicago Board of Trade (CBOT) has active futures markets for 2-, 5-, 10- and 30-year US Treasuries. Table 1 briefly describes the CBOT contracts and pricing conventions.

Like other exchange-traded derivatives, Treasury futures have two advantages: trading is highly liquid and marking-to-market minimizes counterparty risk. The CBOT open auction trading hours are 7:20 am to 2:00 pm, Central Time, Monday through Friday; the CBOT electronic market functions from 6:00 pm to 4:00 pm, Central Time, Sunday through Friday. All Treasury contracts have a March–June-September–December cycle.

A variety of Treasury instruments meet the criteria to be deliverable issues. Table 1 describes the pricing conventions and the characteristics of the assets that may be delivered to satisfy the contracts. The CBOT defines "conversion factors" that adjust the quoted futures prices for the asset that is actually delivered. Despite these conversion factors, one issue will be the "cheapest to deliver". Cash prices at delivery depend on both the conversion factor for a particular bond and the interest accrued on that bond since the last coupon payment.

Although agents frequently use the futures markets for hedging or taking positions on future price movements, only a modest amount of microstructure research has focused on futures markets. Brandt, Kavajecz, and Underwood [11] show that futures and spot market order flow are useful in predicting daily returns in each market and that the type of trader influences the effect of order flow. Mizrach and Neely [59] show that futures markets contribute a substantial amount of price discovery to US Treasury markets. Campbell and Hendry [12] compare price discovery in the 10-year bond and futures contracts in both the United States and Canada.

Seasonality and Announcement Effects

Seasonality and announcement effects are intimately related to the microstructure literature in that the latter seeks to explain how markets with heterogeneous agents react to the release of information.

Seasonality and Macroeconomic Announcements

The earliest studies considered the issue of daily seasonality in Treasuries. Flannery and Protopapadakis [27] docu-

Treasury Market, Microstructure of the U.S., Table 1
Contract Details from the CBOT Treasury Market

Contract	Quote convention	Pricing example	Deliverable asset characteristics
2-year	1/32 and quarters of 32nds	$95 - 060 = 95 + 6/32$ $95 - 062 = 95 + 6.25/32$ $95 - 065 = 95 + 6.5/32$ $95 - 067 = 95 + 6.75/32$	US Treasury notes with a face value ≥ \$200,000 and original maturity ≤ 5 years and 3 months and remaining maturity ≥ 1 year and 9 months from the first day of the delivery month and and remaining maturity ≤ than 2 years from the last day of the delivery month.
5-year	1/32 and halves of 32nds	$90 - 170 = 90 + 17/32$ $90 - 175 = 90 + 17.5/32$	US Treasury notes with a face value ≥ \$100,000 and original maturity ≤ 5 years and 3 months and remaining maturity ≥ 4 year and 2 months from the first day of the delivery month
10-year	1/32 and halves of 32nds	$90 - 170 = 90 + 17/32$ $90 - 175 = 90 + 17.5/32$	US Treasury notes with a face value ≥ \$100,000 and remaining maturity ≤ 10 years remaining maturity ≥ 6 year and 6 months from the first day of the delivery month
30-year	1/32nds	$85 - 12 = 85 + 12/32$	US Treasury bonds with a face value ≥ \$100,000 and if callable: Not callable for at least 15 years from the first day of the delivery month; if not callable: Remaining maturity ≥ 15 years from the first day of the delivery month.

ment differing day-of-the-week patterns in Treasuries and stock indices. The patterns in the prices of Treasuries securities vary by maturity and differ from those found in stock indices. They conclude that no single factor explains seasonal patterns across asset classes. In contrast to this day-of-the-week effect in spot T-bills, Johnston et al. [50] find day-of-the-week effects in government national mortgage association (GNMA) securities, T-note, and T-bond futures, but not in T-bill futures. The fact that day-of-the-week effects exist in spot T-bills but not in T-bill futures points up the importance of futures settlement rules.

Later studies began to consider the effects of macro announcements on price changes, volatility, volume and spreads. Macroeconomic announcements have been an especially popular subject of study because they occur at regular intervals that can be anticipated by market participants. The existence of survey expectations about upcoming macro announcements permits researchers to identify the "shock" component of the announcement, which allows them to investigate the differential effects of anticipated and unanticipated news releases of different magnitudes.

Ederington and Lee [20,21] did the seminal modern work with intraday data on macro announcement effects in bond markets. They found that volatility increases before the announcement and remains elevated for some time afterwards. The employment, PPI, CPI and durable goods orders releases produce the greatest impact of the 9 significant announcements, out of 16 studied. Ederington and Lee [22] follow up on their earlier studies by linking the literatures on seasonality and announcements in the bond market. Comparing the contributions of past volatility, seasonality and announcements in predicting intraday volatility bond futures data and exchange rates, these authors argue that announcements account for much of the apparent seasonality in interest rate volatility.

One of the earliest important results was that bond market prices react more strongly to macro announcements than do equity markets. Fleming and Remolona [32, 34] examined the 25 largest price changes in the GovPX data and related them all to macroeconomic announcements. Fleming and Remolona [34] note: "In contrast to stock prices, US Treasury security prices largely react to the arrival of public information on the economy". Fleming and Remolona [32,33] attribute the relative sensitivity of bond markets to the fact that bond prices depend only on expected discount rates while stock prices are also determined by future expected dividends. Macro announcements can have little or no effect on stock prices if their effects on expected dividends and discount rates offset each other.

Several studies used more sophisticated econometric procedures to evaluate the impact of announcements on persistence in volatility in a full model. Jones, Lamont and Lumsdaine [49] examine volatility patterns in the 5-year Treasury market around US announcements. Daily volatility from an ARCH-M does not persist for days after announcements and the authors interpret this as indicating that agents rapidly incorporate announcement information into prices. Weekly volatility displays a U-shaped pattern; the largest price changes occur on Mondays and Fridays. Further, Jones, Lamont and Lumsdaine [49] find a risk premium in returns on days of announcements. Bollerslev, Cai, and Song [8] also consider the interaction of announcements and persistence in volatility with 5-minute US Treasury bond data. Modeling the intraday volatility patterns and accounting for announcements reveals long-memory in bond market volatility.

An important issue in microstructure is the determination of bid-ask spreads. Balduzzi, Elton, and Green [4] use intraday GovPX data to look at the effects of macro announcements on volume, prices and spreads. Confirming previous findings, prices adjust to news within one minute while increases in volatility and volume persist for up to 60 minutes. Spreads initially widen but then return to normal after 5 to 15 minutes. News releases explain a substantial amount of bond market volatility. Importantly, Balduzzi, Elton, and Green [4] argue that the differential impact of news on long and short bond prices indicates that at least two factors will be needed for models of the yield curve. They also present evidence that discontinuities (jumps) will be important in modeling bond prices.

Some recent papers have relaxed the restrictive assumption that announcements influence Treasury market variables in a linear, symmetric fashion. For example, Christie–David, Chaudhry, and Lindley [15] allow the effects of announcement shocks to depend on the size and sign of the shock. They measure these nonlinear effects on the intraday 10- and 30-year Treasury futures from 1992 to 1996.

Most studies of the effects of volatility have measured such variation with some function of squared returns. One can use the volatility implied by options prices, however, to measure expected volatility over longer horizons. Heuson and Su [45], for example, show that implied volatilities from options on Treasuries rise prior to macro announcements and that volatilities quickly return to normal levels after announcements. Beber and Brandt [6] use intraday, tick data from 1995 to 1999 to determine that macro announcements reduce the variance of the option-implied distribution of US Treasury bond prices. The content of the news and economic conditions explain these changes in higher-order moments. The study attributes the results to time-varying risk premia rather than relative mispricing or changing beliefs.

In a comprehensive study of the impact of US macroeconomic announcements across asset markets, Andersen, Bollerslev, Diebold and Vega [3] study the reaction of international equity, bond and foreign exchange markets. They confirm that US macroeconomic news drives bond prices, as well as those of the other assets.

Monetary Policy Announcements

Researchers have carefully investigated the effects of the Federal Reserve's actions on the Treasury market. While the literature has examined the effect of a wide variety of monetary policy behavior and communications – e. g., open market operations, FOMC news releases, speeches, etc. – on many aspects of Treasury market behavior, a large subset of these papers deal with one specific topic: The effect of federal funds target changes on the Treasury yield curve.

Federal Funds Target Changes and the Treasury Yield Curve The "expectations hypothesis of the term structure" motivates research on how the short- and long-end of the Treasury yield curve react to unexpected changes in the federal funds target rate. That is, if the FOMC increases overnight interest rates, how does this change short- and long-term rates?

Using data on 75 changes in the federal funds target from September 1974 through September 1979, Cook and Hahn [16] find that these target changes caused larger movements in short-term rates than in intermediate- and long-term Treasury rates. A difficulty with interpreting the Cook and Hahn [16] results is that efficient markets presumably can often anticipate most or all of a target change and such expectations are already incorporated into the yield curve. To confront this problem, Kuttner [53] decomposes target changes into anticipated and unanticipated components, finding –unsurprisingly – that Treasury rates respond much more strongly to unanticipated changes and that the results are consistent with the expectations hypothesis of the term structure. That is, the anticipated component of an interest rate change does not affect expectations. Hamilton [41] carefully reexamines the work of Kuttner [53], showing that it is robust to uncertainty about the dates of target changes and the effect of learning by market participants.

Poole and Rasche [62] also decompose federal funds target changes into expected and unexpected components – but use a later contract month than Kuttner [53] to avoid problems associated with computation of the contract payoff. They find that interest rates across the maturity spectrum fail to respond to the anticipated components of the changes in the intended funds rate.

Poole, Rasche and Thornton [63] consider how changes in FOMC procedures affect the impact of target changes on interest rates. This study first succinctly describes the changes in FOMC procedures in the 1990s. The FOMC began to contemporaneously announce policy actions in 1994 and adopted this as formal policy in 1995. Starting in August 1997, each policy directive has included the quantitative value of the "intended federal funds rate". And since 1999, the FOMC has issued a press release after each meeting with the value for the "intended federal funds rate" and, in most cases, an assessment of the balance of risks. After describing such procedural changes, Poole, Rasche and Thornton [63] go on to consider the response of the Treasury yield curve to funds rate target changes both before and after the FOMC began contemporaneously announcing target changes in 1994. In doing so, these authors account for measurement error in expectations and uncertainty about the dates of target changes and even whether market participants understood that the Federal Reserve was targeting the funds rate prior to 1994. They assess the market's knowledge of targeting by examining news reports. While short-rates respond similarly in both subperiods, long rates do not respond as strongly to funds rate target changes after 1994. The authors interpret their results as being consistent with the Fed's greater transparency about long-run policy in the second subsample. With long-run expectations more firmly anchored, unexpected changes in the funds target have smaller effects on long rates.

One puzzle that has emerged from this literature is that the average effect of changes in the federal funds target on the yield curve is modest, despite the facts that such changes should be an important determinant of the yield curve and that yields are highly volatile around FOMC announcements. Fleming and Piazzesi [31] claim to partially resolve this puzzle by illustrating that such yield changes depend on the shape of the yield curve.

This literature on the reaction of the Treasury market to monetary policy has become progressively more sophisticated in assessing market expectations of Fed policy and modeling institutional features of the futures market and Fed operations. Nevertheless, the underlying conclusion that unanticipated target changes lead to large price increases on short-term Treasuries and smaller changes on the prices of long-term Treasuries has been remarkably robust.

Other Federal Reserve Behavior and the Treasury Market There has been a substantial literature analyzing how other types of Federal Reserve behavior have influenced the Treasury market. The literature has considered open market operations, FOMC statements, Congressional testimonies, and FOMC member speeches.

Open market operations are similar to macroeconomic announcements in that they are potentially important bond market events, occurring at regularly scheduled times. Harvey and Huang [43] used intraday data from 1982 to 1988 to examine how Federal Reserve open market operations influenced foreign exchange and bond markets. The paper finds that Treasury market volatility increases during open market operations, irrespective of whether they add or drain reserves. Oddly, volatility increases even more during the usual time for open market operations if there are no such transactions. The authors interpret this finding as indicating that open market operations actually smooth volatility.

Early studies made the simplifying assumption that the effect of macro announcements on the Treasury market was constant over time. This is not necessarily the case, of course. For example, the effect of macro announcements on the Treasury market might depend on monetary policy priorities. Kearney [52] characterizes the changing response of daily 3-month Treasury futures to the employment report over 1977 to 1997 and relates it to the changing importance of employment in the Fed's reaction function.

de Goeij and Marquering [17] also considers how both macro announcements and monetary policy events affect the US Treasury market. Using daily data from 1982 to 2004 de Goeij and Marquering [17] find that macro news announcements strongly affect the daily volatility of longer-term Treasury instruments while FOMC events affect the volatility of shorter-term instruments.

Some studies have explored more esoteric components of information about monetary policy. Boukus and Rosenberg [9], for example, use Latent Semantic Analysis to decompose the information content of FOMC minutes from 1987 to 2005. They then relate the information content to current and future economic conditions. Chirinko and Curran [13] argue that Federal Reserve speeches, testimonies, and meetings increase price and trading volatility on the 30-year bond market. FOMC meetings are the most important of the events considered. They go on to consider whether these Federal Reserve events merely create noise or transmit information about the future policy decisions or the state of the economy. They conclude that such events may reduce welfare by "overwhelming private information", creating herding behavior.

Announcements and Liquidity Variation

The literature on variation in liquidity and price effects overlaps with the literature on macroeconomic announcements. The seminal work of Amihud and Mendelson [2] showed that yields on short-time-to-maturity Treasuries vary inversely with liquidity. That is, more liquid assets have lower yields/higher prices. Harvey and Huang [43] discovered elevated volatility in interest rate (and foreign exchange) futures markets, in the first 60–70 minutes of trading on Thursdays and Fridays. Ederington and Lee [20] confirmed Harvey and Huang [43]'s speculation that major macroeconomic announcements – especially the employment report, the PPI, the CPI, and durable goods orders – create the intraday and intraweek patterns in the volatility of Treasury bond futures. Volatility is very high after announcements and remains elevated for hours. Fleming and Remolona [32] extend this work to show that the 25 greatest surges in activity in the 5-year on-the-run bond market came on macroeconomic announcement days, within 70 minutes of the announcement. The most important announcements for trading surges were employment reports, fed funds targets, 30-year auctions, 10-year auctions, the CPI, NAPM surveys, GDP, retail sales, and 3-year auctions. Releases that affect prices also matter for trading activity. Fleming and Remolona [32] observe that timeliness, the degree of surprise in the announcement and market uncertainty also increase announcements' impact on trading.

Researchers continued to explore the impact of variation in liquidity caused by other events. For example, Fleming [28] exploits exogenous variation in Treasury issuance to show that securities that are "reopened" – the Treasury sells additional quantities of existing securities – have greater liquidity, lower spreads, than comparable assets. Paradoxically, this higher liquidity does not produce lower yields for the reopened securities.

More recent papers have explored variation in liquidity and volatility across markets. Chordia, Sarkar and Subrahmanyam [14] estimate a vector autoregression (VAR) in liquidity and volatility variables in stock and bond markets. They find that common factors make the variables' innovations highly correlated. Volatility shocks predict liquidity variables.

End-of-the-Year Patterns in One-Month Treasury Bills

The previous sets of papers studied daily and intraday seasonality, often as caused by macroeconomic or Federal Reserve announcements. Short-term Treasury bills also exhibit year-end seasonality, however. Market participants consider Treasury market instruments of 30 days or less to be highly liquid, close – but not perfect – substitutes for cash. The fact that short-term Treasuries are not perfect substitutes for cash is presumably what allows the New York Desk to use open market operations to manipulate short-term interest rates through a liquidity effect. A peculiar year-end pattern in one-month Treasury yields reinforces this evidence that such Treasuries are not perfect substitutes for cash.

Following on related work of Griffiths and Winters [40] in repos, Griffiths and Winters [39] find that yields on one month T-Bills (and other one-month securities) increase significantly at the beginning of December, remain high during December, and return to normal a few days before the year-end. This pattern does not exist in three-month T-bills. Neely and Winters [60] find similar patterns in the one-month LIBOR futures market.

Griffiths and Winters [38,39,40] explain this December effect by asserting that a year-end preference for liquidity drives the year-end surge in short-term interest rates. Debt holder (lenders in the money markets) start to liquidate their one-month securities in the last few days of November to meet cash obligations at the end of December. This preference for liquidity drives up one-month interest rates for most of December. Liquidity demand returns to normal at the end of December as investors repurchase short-term instruments, and interest rates return to normal levels.

Discontinuities in the US Treasury Market

The literature on discontinuities (or jumps) in Treasury prices is closely related to the literature on announcements, as announcements are obvious candidates to explain jumps. Three recent papers have looked at discontinuities in US Treasury prices. Huang [47] estimates daily jumps with bi-power variation on 10 years of 5-minute data on S&P 500 and US T-bond futures to measure the response of volatility and jumps to macro news. He identifies a major role for payroll news in bond market jumps by analyzing their conditional distributions and regressing continuous and jump components on measures of disagreement and uncertainty concerning future macroeconomic states. Huang [47] also finds that the bond market is relatively more responsive than the equity market.

Dungey, McKenzie, and Smith [19] estimate jumps and cojumps (simultaneous discontinuities in multiple markets) in the term structure of US Treasury rates. They find that the middle of the yield curve often cojumps with one of the ends, while the ends of the curve exhibit a greater tendency for idiosyncratic jumps. Macro news is strongly associated with cojumps in the term structure. Using BrokerTec data from 2003–2005, Jiang, Lo, and Verdelhan [48] extend this work by focusing on the role of liquidity shocks – estimated from the limit order book – in jumps and the relation of jumps to order flow and price discovery.

Lahaye, Laurent and Neely [54] examine jumps and cojumps across foreign exchange, stock, gold and 30-year Treasury futures. Discontinuities in bond futures prices were larger but less frequent than those in foreign exchange rates and smaller and about as frequent as those in equity markets. News announcements appear to cause many cojumps of bond prices with prices of other types of assets.

Order Flow in the US Treasury Market

The effect of order flow on prices has been a popular recent topic in microstructure. Several papers have explored the impact of order flow on prices and the ways in which macro/monetary announcements influence these impacts.

Huang, Cai, and Wang [46] use intraday 1998 GovPX spot data on the 5-year Treasury note to characterize trading patterns of primary dealers, announcement effects and volatility-volume relations. The paper finds that both public information (i. e., announcements) and dealer inventory/order flow affect trading frequency.

Green [37] uses the Madhavan, Richardson, and Roomans [55] model to study the impact of GovPX trading in 5-year around announcements. Order flow has its largest price impact after large macro surprises, times of greater uncertainty about the announcement, and times of high liquidity. Green [37] concludes that order flow does reveal information about riskless rates.

Brandt and Kavajecz [10] find that order flow imbalances can explain up to 26% of the day-to-day variation in yields on non-announcement days. In contrast to Green [37], they find that order flow has its strongest impact at times of low liquidity. Brandt, Kavacejz, and Underwood [11] extend the work of Brandt and Kavajecz [10] to control for trader type and macroeconomic announcements in explaining the impact of bond market order flow on futures prices.

Menkveld, Sarkar, and Van der Wel [56] confirm earlier conclusions that announcements have significant effects on 30-year Treasury yields and they also find that customer order flow is much more informative on announcement days than on non-announcement days. They go on to investigate the profits that different types of traders make on announcement and non-announcement days.

At high frequencies, order flow is highly autocorrelated. A dynamic analysis of the market resilience requires modeling this formally. We turn to empirical modeling of the Treasury market order book in the next section.

Modeling the Limit Order Book

A purchase or a sale of a Treasury bond influences prices directly as trades work their way up the supply or demand curves. We would like to know whether these effects are large and long-lasting. To address this question, we must introduce a dynamic model of the limit order book.

Hasbrouck [44] proposed to study intra-day price formation with a standard bivariate vector autoregressive (VAR) model. Time t here is measured in 1-minute intervals. Let r_t be the percentage change in the transaction price and x_t^0 be the sum of signed trade indicators (+1 for buyer initiated, -1 for seller initiated) over minute t. Treasury market data sets typically indicate trade initiation as a "hit" -1 or a "take" +1.

The bivariate vector autoregression assumes that causality flows from trade initiation to returns by permitting r_t to depend on the contemporaneous value for x_t^0, but not allowing x_t^0 to depend on contemporaneous r_t. The model for returns is specified as follows

$$\begin{bmatrix} r_t \\ x_t^0 \end{bmatrix} = \sum\nolimits_{i=1}^{5} \begin{bmatrix} a_{r,i} \\ a_{x,i} \end{bmatrix} r_{t-i} + \begin{bmatrix} \sum_{i=0}^{15} b_{r,i} \\ \sum_{i=1}^{15} b_{x,i} \end{bmatrix} x_{t-i}^0 + \begin{bmatrix} u_{r,t} \\ u_{x,t} \end{bmatrix} . \quad (1)$$

Mizrach and Neely [58] use 5 lags of the return series and 15 lags of the signed trades. The market impact is then defined as the dynamic effect of a buy shock to the return series,

$$\frac{\partial r_{t+n}}{\partial x_t} . \quad (2)$$

Mizrach and Neely [58] provide 15 minute market impact estimates from the GovPX market in 1999. The 2-year note is most resilient with prices only 0.0042% higher following a buyer initiated trade. The 30-year bond is the least liquid, with prices rising 0.0229% following a buy order. Mizrach and Neely also report 2004 estimates for the Cantor electronic limit order book. Market impacts range from 45 to 88% lower in the more liquid eSpeed ECN market. Fleming and Mizrach [30] find further reductions in market impacts on the BrokerTec ECN for 2005 and 2006.

Price Discovery

A crucial issue in the market microstructure literature is *price discovery*. This is the process by which prices embed new information. In the Treasury market, price discovery occurs in both the secondary spot market and in the futures markets at the Chicago Board of Trade (CBOT). The degree to which each market contributes to price discovery is a natural issue to address.

To investigate relative price discovery in these two Treasury markets, Mizrach and Neely [59] follow Hasbrouck [44] and assume that the price series have a unit root, are cointegrated, and have an r^{th} order VAR representation,

$$p_t = \Phi_1 p_{t-1} + \Phi_2 p_{t-2} + \cdots + \Phi_r p_{t-r} + u_t .$$

It follows that the N returns,

$$r_t = \begin{bmatrix} p_{1,t} - p_{1,t-1} \\ \vdots \\ p_{N,t} - p_{N,t-1} \end{bmatrix} = \Delta p_t , \quad (3)$$

have the convenient Engle–Granger [24] error-correction representation,

$$\Delta p_t = \alpha z_{t-1} + A_1 \Delta p_{t-1} + \cdots + A_r \Delta p_{t-r-1} + u_t , \quad (4)$$

where z_t is an error-correction term of rank $N-1$.

We analyze price discovery using the moving average representation of our return process (3),

$$\Delta p_t = \Theta(L)\varepsilon_t . \quad (5)$$

The disturbances are mean zero and serially uncorrelated, $E[\varepsilon_{i,t}] = 0$ and $\text{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-r}] = 0$, but they may be contemporaneously correlated, $\text{cov}[\varepsilon_{i,t}, \varepsilon_{j,t}] \neq 0$.

The information share is related to the long run impulse responses, $\Theta(1) = \sum_{j=0}^{\infty} \Theta(L^j)$, the permanent effect of the shock vector on the Treasury prices. Cointegration makes the long run multipliers common across all markets,

$$\Theta(1) = \begin{bmatrix} \theta_1 & \cdots & \theta_N \\ \vdots & & \vdots \\ \theta_1 & \cdots & \theta_N \end{bmatrix} . \quad (6)$$

To eliminate contemporaneous correlation among the error terms in (5), we decompose $\Omega = E\left[\varepsilon_t \varepsilon_t'\right]$, the $N \times N$ covariance matrix, to find a lower triangular matrix M, whose i, jth element we denote m_{ij}, such that $MM' = \Omega$. The Hasbrouck [44] information share for market j is defined as

$$H_j = \frac{\left[\sum_{i=j}^{n} \theta_i m_{ij}\right]^2}{\left[\sum_{i=1}^{n} \theta_i m_{i1}\right]^2 + \left[\sum_{i=2}^{n} \theta_i m_{i2}\right]^2 + \cdots + (\theta_n m_{nn})^2} , \quad (7)$$

where the θ_is are the elements of row i of the long-run multipliers in (6). Because the Choleski decomposition is not unique, the information share will vary with the order of the equations in the VAR.

Mizrach and Neely [59] pair spot and maturity matched futures for the 2-year, 5-year and 10-year on-the-run spot notes. This calculation requires us to adjust futures prices according to the on-the-run spot instruments with which we compare them. The CBOT provides adjustment factors for each instrument. These adjustments typically make a single bond the cheapest to deliver (CTD), but the CTD is typically off-the-run. Nevertheless, the CTD off-the-run bonds and the most liquid on-the-run bonds are very close substitutes – their daily returns are highly correlated – so it is reasonable to examine price discovery between futures prices and on-the-run bonds, despite the fact that they are not identical.

Mizrach and Neely [59] find that information shares rise with the growth of the GovPX market, but fall as the ECNs take market share from GovPX voice markets. The spot market share is highest for the 2-year note, reaching 86%, while the 10-year spot market share never exceeds 50%. In addition, relative market liquidity measures like spreads, trades and volatility each strongly explain daily relative price discovery shares. Mizrach and Neely [59] compute both upper and lower bound estimates of the information shares. They also report estimates based on the Harris, McInish and Wood [42] methodology.

Campbell and Hendry [12] find similar results for the Canadian government bond market. They find that the information share in the 10-year spot note is below 50% in nearly all their sample of several months between 2002 and 2004. Upper and Werner [67] find that price discovery in the German Bund is dominated by the futures market, and in times of stress, like the 1998 Long Term Capital Management Crisis, the spot market information share falls to essentially zero. Upper and Werner [67], however, compare the futures market to the relatively illiquid, CTD bonds. This might explain their finding that the spot market does very little price discovery.

Future Directions

This article has reviewed the microstructure of the US Treasury market. The Open Market Desk at the Federal Reserve Bank of New York plays a uniquely important role in the Treasury market by using transactions in those securities to adjust the level of bank reserves. Primary dealers are key players in both Treasury auctions and the Fed's open market operations. The Treasury market consists of several phases: when-issued, primary, on-the-run and off-the-run. Two ECNs, eSpeed and BrokerTec, intermediate the most active trading, during the on-the-run phase. The Treasury futures market at the CBOT complements trading in the spot market.

Treasury markets exhibit end-of-year, daily and intraday seasonality. Macro and Federal Reserve announcements are responsible for a substantial part of the daily and intraday seasonality. The literature studying the impact of order flows on Treasury prices has also considered how macro news and Federal Reserve actions influence such impact.

The futures markets in Chicago play an important role in price discovery, and a discussion of Treasury microstructure needs to take this into account. Both spot and futures markets are quite resilient and recent research on the Treasury ECNs suggest that the market continues to become more liquid. Fleming and Mizrach [30] report that volume has increased almost 5 times since 2001. This increase in trading volume accompanies a decline in the importance of the primary dealers. The Financial Times reported in March 2007 that hedge funds accounted for 80% of trading activity in the Treasury market with only a 20% share for the primary dealers. One large fund alone, Citadel, accounts for 10% of the trading volume on eSpeed and BrokerTec. It was perhaps inevitable that trading by the millisecond would come to the Treasury market as it did to equities and foreign exchange. Perhaps we should only be surprised that it took so long.

The Treasury market plays a central role in the credit market. Times of financial crisis highlight the Treasury market's role as a safe haven for investors both in the US and overseas. Treasury securities also serve as benchmarks for complex derivatives like mortgage backed securities and structured loans like collateralized debt obligations. The microstructure of the US Treasury market is fundamental to our understanding of the global financial markets.

Bibliography

1. Akhtar MA (1997) Understanding open market operations. Public Information Department, Federal Reserve Bank of New York, New York
2. Amihud Y, Mendelson H (1991) Liquidity, maturity, and the yields on US treasury securities. J Finance 46:1411–25
3. Andersen TG, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond and foreign exchange markets. J Int Econ 73:251–77
4. Balduzzi P, Elton EJ, Green TC (2001) Economic news and bond prices: Evidence from the US treasury market. J Financial Quant Anal 36:523–43
5. Barclay MJ, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from treasuries going off the run. J Finance 61:2395–2414

6. Beber A, Brandt MW (2006) The effect of macroeconomic news on beliefs and preferences: Evidence from the options market. J Monetary Econ 53:1997–2039
7. Bernanke BS (2005) Implementing monetary policy, remarks at the redefining investment strategy education symposium, Dayton. This is a public speech by Bernanke http://www.federalreserve.gov/boarddocs/speeches/2005/20050330/default.htm
8. Bollerslev T, Cai J, Song FM (2000) Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the us treasury bond market. J Empir Finance 7:37–55
9. Boukus E, Rosenberg JV (2006) The information content of FOMC minutes. Working paper, Federal Reserve Bank of New York, New York
10. Brandt MW, Kavajecz KA (2004) Price discovery in the us treasury market: The impact of orderflow and liquidity on the yield curve. J Finance 59:2623–2654
11. Brandt MW, Kavajecz KA, Underwood SE (2007) Price discovery in the treasury futures market. J Futur Mark 27:1021–1051
12. Campbell B, Hendry S (2007) Price discovery in canadian and US 10-year government bond markets. Working Paper 07–43, Bank of Canada, Ottawa
13. Chirinko RS, Curran C (2006) Greenspan shrugs: Formal pronouncements, bond market volatility, and central bank communication. Presented at: The American Economic Association Meetings
14. Chordia T, Sarkar A, Subrahmanyam A (2005) An empirical analysis of stock and bond market liquidity. Rev Financial Stud 18:85–129
15. Christie-David R, Chaudhry M, Lindley JT (2003) The effects of unanticipated macroeconomic news on debt markets. J Financial Res 26:319–39
16. Cook T, Hahn T (1989) The effect of changes in the federal funds rate target on market interest rates in the 1970s. J Monetary Econ 24:331–351
17. de Goeij P, Marquering W (2006) Macroeconomic announcements and asymmetric volatility in bond returns. J Bank Finance 30:2659–2680
18. Duffie D (1996) Special repo rates. J Finance 51:493–526
19. Dungey M, McKenzie M, Smith V (2007) News, no-news and jumps in the us treasury market. Cambridge University (unpubl)
20. Ederington LH, Lee JH (1993) How markets process information: News releases and volatility. J Finance 48:1161–91
21. Ederington LH, Lee JH (1995) The short-run dynamics of the price adjustment to new information. J Financial Quant Anal 30:117–34
22. Ederington LH, Lee JH (2001) Intraday volatility in interest-rate and foreign-exchange markets: ARCH, announcement, and seasonality effects. J Futur Mark 21:517–52
23. Elton EJ, Green TC (1998) Tax and liquidity effects in pricing government bonds. J Finance 53:1533–1562
24. Engle R, Granger C (1987) Co-integration and error correction representation, estimation and testing. Econometrica 55:251–276
25. Fabozzi FJ, Fleming MJ (2005) US treasury and agency securities. In: Fabozzi FJ (ed) The Handbook of Fixed Income Securities, 7th edn. McGraw Hill, New York, pp 229–250
26. Federal Reserve Bank of New York, Markets Group (2007) Domestic open market operations during 2006. Federal Reserve Bank of New York. http://www.ny.frb.org/markets/omo/omo2006.pdf
27. Flannery M, Protopapadakis A (1988) From T-bills to common stocks: Investigating the generality of intra-week return seasonalities. J Finance 43:431–450
28. Fleming M (2002) Are larger treasury issues more liquid? Evidence from bill reopenings. J Money Credit Bank 34:707–735
29. Fleming M (2003) Measuring treasury market liquidity. Fed Reserv Bank New York Econ Policy Rev 9:83–108
30. Fleming M, Mizrach B (2008) The microstructure of a US treasury ECN: The BrokerTec platform. Working paper. Federal Reserve Bank of New York, New York
31. Fleming M, Piazzesi M (2005) Monetary policy tick-by-tick. Working paper No ID. Federal Reserve Bank of New York, New York
32. Fleming M, Remolona EM (1997) What moves the bond market? Fed Reserv Bank New York Econ Policy Rev 3:31–50
33. Fleming M, Remolona EM (1999) Price formation and liquidity in the US treasury market: The response to public information. J Finance 54:1901–1915
34. Fleming M, Remolona EM (1999) What moves bond prices? J Portf Manag 25:28–38
35. Garbade KD, Ingber JF (2005) The treasury auction process: Objectives, structure, and recent adaptations. Fed Reserv Bank New York Curr Issues Econ Finance 11:1–11
36. Goldreich D, Hanke B, Nath P (2005) The price of future liquidity: Time-varying liquidity in the US treasury market. Rev Finance 9:1–32
37. Green TC (2004) Economic news and the impact of trading on bond prices. J Finance 59:1201–1233
38. Griffiths M, Winters D (1997) On a preferred habitat for liquidity at the turn-of-the-year: Evidence from the term-repo market. J Financial Serv Res 12:21–38
39. Griffiths M, Winters D (2005) The turn-of-the-year in money markets: Tests of risk-shifting window dressing and preferred habitat hypotheses. J Bus 78:1337–1364
40. Griffiths M, Winters D (2005) The year-end price of risk in a market for liquidity. J Invest Manag 3:99–109
41. Hamilton JD (2008) Assessing monetary policy effects using daily fed funds futures contracts. Fed Reserv Bank St Louis Rev 90:377–393
42. Harris F, McInish T, Wood R (2002) Security price adjustment across exchanges: An investigation of common factor components for dow stocks. J Financial Mark 5:277–308
43. Harvey CR, Huang RD (2002) The impact of the federal reserve bank's open market operations. J Financial Mark 5:223–57
44. Hasbrouck J (1991) Measuring the information content of stock trades. J Finance 46:179–207
45. Heuson AJ, Su T (2003) Intra-day behavior of treasury sector index option implied volatilities around macroeconomic announcements. Financial Rev 38:161–77
46. Huang RD, Cai J, Wang X (2002) Information-based trading in the treasury note interdealer broker market. J Financial Intermed 11:269–296
47. Huang X (2006) Macroeconomic news announcements, financial market volatility and jumps. Duke University (unpubl)
48. Jiang GJ, Lo I, Verdelhan A (2007) Why do bond prices jump? A study of the US treasury market. Eller College of Management, University of Arizona (unpubl)

49. Jones CM, Lamont O, Lumsdaine RL (1998) Macroeconomic news and bond market volatility. J Financial Econ 47:315–337
50. Johnston E, Kracaw W, McConnell J (1991) Day-of-the-week effects in financial futures: An analysis of GNMA, T-bond, T-note, and T-bill contracts. J Financial Quant Anal 26:23–44
51. Kamara A (1994) Liquidity, taxes, and short-term treasury yields. J Financial Quant Anal 29:403–417
52. Kearney AA (2004) The changing impact of employment announcements on interest rates. J Econ Bus 54:415–429
53. Kuttner KN (2001) Monetary policy surprises and interest rates: Evidence from the fed funds futures market. J Monetary Econ 47:523–544
54. Lahaye J, Laurent S, Neely CJ (2007) Jumps, cojumps and macro announcements. Working Paper 2007–032A, Federal Reserve Bank of St. Louis, St. Louis
55. Madhavan A, Richardson M, Roomans M (1997) Why do securities prices change? A transaction-level analysis of NYSE stocks. Rev Financial Stud 10:1035–1064
56. Menkveld AJ, Sarkar A, Van der Wel M (2006) Customer flow, intermediaries, and the discovery of the equilibrium riskfree rate. Working paper. Federal Reserve Bank of New York, New York
57. Meulendyke AM (1998) US monetary policy & financial markets. Federal Reserve Bank of New York, New York
58. Mizrach B, Neely CJ (2006) The transition to electronic communication networks in the secondary treasury market. Fed Reserv Bank St. Louis Rev 88:527–541
59. Mizrach B, Neely CJ (2008) Information shares in the US treasury market. J Bank Finance 32:1221–1233
60. Neely CJ, Winters DB (2006) Year-end seasonality in one-month LIBOR derivatives. J Derivatives 13:47–65
61. Nyborg KG, Sundaresan S (1986) Discriminatory versus uniform treasury auctions: Evidence from when-issued transactions. J Financial Econ 42:63–104
62. Poole W, Rasche RH (2000) Perfecting the market's knowledge of monetary policy. J Financial Serv Res 18:255–298
63. Poole W, Rasche RH, Thornton DL (2002) Market anticipations of monetary policy actions. Fed Reserv Bank St. Louis Rev 84:65–94
64. Seligman J (2006) Does urgency affect price at market? An analysis of US treasury short term finance. J Money Credit Bank 38:989–1012
65. Simon DP (1991) Segmentation in the treasury bill market: Evidence from cash management bills. J Financial Quant Anal 26:97–108
66. Simon DP (1994) Further evidence on segmentation in the treasury bill market. J Bank Finance 18:139–151
67. Upper C, Werner T (2002) Tail wags dog? Time-varying information shares in the bund market. Working Paper 24/02, Bundesbank, Frankfurt
68. Vayanos D, Weill P (2008) A search-based theory of the on-the-run phenomenon. J Finance 63:1361–1398
69. Warga A (1992) Bond returns, liquidity, and missing data. J Financial Quant Anal 27:605–17

List of Glossary Terms

Index