INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE

A Long View of Research and Practice in Operations Research and Management Science

The Past and the Future

ManMohan S. Sodhi
Christopher S. Tang
Editors

Springer

International Series in Operations Research & Management Science

Volume 148

For further volumes:
http://www.springer.com/series/6161

ManMohan S. Sodhi · Christopher S. Tang
Editors

A Long View of Research and Practice in Operations Research and Management Science

The Past and the Future

Springer

Editors
ManMohan S. Sodhi
City University
Cass Business School
Bunhill Row 106
EC1Y 8TZ London
United Kingdom
m.sodhi@city.ac.uk

Christopher S. Tang
University of California
Los Angeles
Anderson School of Management
Westwood Plaza 110
90095 Los Angeles California
Box 951481
USA
ctang@anderson.ucla.edu

ISSN 0884-8289
ISBN 978-1-4419-6809-8 e-ISBN 978-1-4419-6810-4
DOI 10.1007/978-1-4419-6810-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010934120

Foreword

As generation of academics and practitioners follows generation, it is worthwhile to compile long views of the research and practice in the past to shed light on research and practice going forward. This collection of peer-reviewed chapters is intended to provide such a long view. The effort is motivated by the views of Professor Arthur M. Geoffrion, who we seek to honor for not only his considerable contribution to OR/MS research in the past decades but also his continuing championship and involvement in matters pertaining to the education and practice of OR/MS.

Professor Geoffrion's contributions are well highlighted in "About Professor Arthur M. Geoffrion," but I would like to add a personal note. When I was an unknown first year assistant professor and Art was an established superstar, he took the trouble to obtain a copy of my thesis, read it, and call me to offer advice and encouragement. His advice covered both high-level direction and important details and was delivered with a charm and humor that made it easy to accept. For example, I was pretty green then as a mathematician and had used the term "cycle-less graph" in my thesis. Art's wry remark was "'Cycle-less graph,' that must be an east coast term. Here in California, and I think most of the world, that's called an 'acyclic graph'." My thesis concerned using Lagrange multipliers to solve job shop scheduling problems. Art subsequently described in the article Geoffrion, AM. (1974) Lagrangean relaxation for integer programming. Math Program Stud 2:82–114 how this work and several other problem-specific uses of Lagrange multipliers could be embraced within a powerful concept he called "Lagrangian Relation."

The target audience of this book is young researchers, graduate/advanced undergraduate students from OR/MS and related fields like computer science, engineering, and management as well as practitioners who want to understand how OR/MS modeling came about over the past few decades and what research topics or modeling approaches they could pursue in research or application.

This book contains a collection of chapters written by leading scholars/practitioners who have continued their efforts in developing and/or implementing innovative OR/MS tools for solving real-world problems. In this book, the contributors share their perspectives about the past, present, and future of OR/MS theoretical development, solution tools, modeling approaches, and applications. Specifically, this book collects chapters that offer insights about the following topics:

- Survey articles taking a long view over the past two or more decades to arrive at the present state of the art while outlining ideas for future research. Surveys focus on use of a particular OR/MS approach, e.g., mathematical programming (LP, MILP, etc.), and solution methods for particular family of application, e.g., distribution system design, distribution planning system, health care.
- Autobiographical or biographical accounts of how particular inventions (e.g., structured modeling) were made. These could include personal experiences in early development of OR/MS and an overview of what has happened since.
- Development of OR/MS mathematical tools (e.g., stochastic programming, optimization theory).
- Development of OR/MS in a particular industry sector such as global supply chain management.
- Modeling systems for OR/MS and their development over time as well as speculation on future development (e.g., LINDO, LINGO, and What's *Best*!).
- New applications of OR/MS models (e.g., happiness).

I believe this book will stimulate others to follow Professor Geoffrion's footsteps in making OR/MS a vibrant community.

The Wharton School, Marshall Fisher
University of Pennsylvania,
Philadelphia, PA, USA
February 2010

Acknowledgments

We would like to thank Professor Fred Hillier (Stanford University), the editor of Springer's International Series in Operations Research and Management Science, who strongly encouraged us to work on this book from the very beginning. The book received strong support from colleagues from many universities and companies, many of them committing to contribute to this collection. We would like to express our sincere appreciation to them for providing their leading edge research for this book.

Name (in alphabetical order)	Affiliation	Chapter
Mustafa Atlihan, Kevin Cunningham, Gautier Laude, Linus Schrage	LINDO Systems, University of Chicago	Challenges in adding a stochastic programming/scenario planning capability to a general purpose optimization modeling system
Manel Baucells, Rakesh Sarin	IESE Business School, University of California, Los Angeles	Optimizing happiness
Dirk Beyer, Scott Clearwater, Kay-Yut Chen, Qi Feng, Bernardo A. Huberman, Shailendra Jain, Alper Sen, Hsiu-Khuern Tang, Zainab Jamal, Bob Tarjan, Krishna Venkatraman Julie Ward, Alex Zhang, Bin Zhang	M-Factor, Inc., Hewlett-Packard Labs, University of Texas at Austin, Bilkent University, Intuit	Advances in business analytics at HP Laboratories

John Birge	University of Chicago	The persistence and effectiveness of large-scale mathematical programming strategies: Projection, outer linearization, and inner linearization
Gerald G. Brown (and Richard E. Rosenthal, deceased)	Naval Postgraduate School	Optimization tradecraft: Hard-won insights from real-world decision support (reprinted with permission from INFORMS)
Daniel Dolk	Naval Postgraduate School	Structured modeling and model management
Donald Erlenkotter	University of California, Los Angeles	Economic planning models for India in the 1960s
Robert Fourer	Northwestern University	Cyber-infrastructure and optimization
Arthur M. Geoffrion, Glenn Graves	University of California, Los Angeles	Multi-commodity distribution system design by Bender's decomposition (reprinted with permission from INFORMS)
Hau L. Lee	Stanford University	Global trade process and supply chain management
Grace Lin, Ko-Yang Wang	World Resource Optimization Inc., IBM Global Business Services	Sustainable globally integrated enterprise
Richard Powers	Formerly at INSIGHT Inc.	Retrospective: 25 years applying management science to logistics
ManMohan S. Sodhi, Christopher S. Tang	City University London, University of California, Los Angeles	Capitalizing on our strengths to avail opportunities in the face of weakness and threats
Mark S. Daskin, Sanjay Mehrotra, Jonathan Turner	Northwestern University, University of Michigan	Perspectives on healthcare resource management problems

Last, but not least, we are grateful to Mirko Janc for typesetting each chapter beautifully and expeditiously. Of course, we are responsible for any errors that may occur in this book as a result of our editing or our own writing.

ManMohan S. Sodhi, London
Christopher S. Tang, Los Angeles

Contents

About Professor Arthur M. Geoffrion

Arthur Geoffrion is the James A. Collins Professor of Management Emeritus (recalled) at the UCLA Anderson School of Management. He received his Ph.D. in operations research from Stanford University in 1965, following B.M.E. and M.I.E. degrees from Cornell University. He has been on the UCLA faculty since that time.

As OR/MS has evolved over time, so has Professor Geoffrion's research. In the late 1960s and early 1970s, he focused on mathematical programming techniques for solving large-scale problems efficiently. These included linearization (cf., [2]), duality (cf., [3]), integer programming (cf., [16]), Lagrangian relaxation (cf., [4]), multi-criterion optimization and decomposition techniques for special structures (cf., [12]). Computational cost at the time was high and computational capability was quite limited relative to what we are used to now and researchers needed to focus on computational efficiency. Geoffrion and Graves [12] presented an efficient solution using Bender's decomposition for solving practical, and therefore large-scale, multi-commodity distribution problems while Geoffrion and Marsten [16] provided a systematic framework for dealing with integer programming problems. Using the newly developed ideas, he helped develop and implement distribution design systems based on mathematical programming at many companies [17] through INSIGHT, Inc., a management consulting firm he co-founded in 1978 that specializes in optimization-based applications in supply-chain management and production planning. He was also consultant to government agencies on applications of optimization to problems of distribution, production, and capital budgeting.

During the 1980s, his interests turned to modeling formalisms and computer-based modeling environments as an approach to improving the quality, productivity, and acceptability of OR/MS in practice. From his experience with companies and government agencies, he realized that there was a need to develop a formalized way to manage models and related data. Managers wanted interface facilities that would "make it easy to build and solve complex models" [18]. In the early 1980s, database management systems were well developed, but there were no unified model management systems to enable users to retrieve or modify models. Without a unified model management system, companies found it difficult to re-use existing models by expanding or otherwise modifying them to meet changing needs. There was thus need for a model management system (MMS) that would (a) have a uniform computer-executable model representation that supports multiple views of a model

as in relational model for database management; (b) support modeling languages; (c) support multiple OR/MS tools (simulations, regressions, queueing, optimization, etc.); and (d) allow separation of models, data, and solvers. Geoffrion offered structured modeling [6–10] with these properties and consequently received much attention by researchers [1].

Since the mid-1990s, his interests have centered on the implications of the Internet and digital economy for management and for management science. By viewing that the "network is the computer," Geoffrion changed his focus to the digital economy. OR/MS can play an important role in the digital economy because OR/MS is equipped to copy large scales of data and complex problems [13]. At the same time, the digital economy can influence the development and deployment of OR/MS. Geoffrion and Krishnan [14, 15] highlight this "mutual impact" in a two-part special issue of *Management Science*.

With over 60 highly cited papers, Professor Geoffrion's research is well recognized (cf., [5, 11]). His research has been supported by about 45 grants and contracts, including many from the National Science Foundation and the Office of Naval Research. His work in the area of distribution planning was awarded a NATO System Science Prize.

His service to the OR/MS community goes well beyond his research. His editorial service includes 8 years as department editor (mathematical programming and networks) of *Management Science*, posts at *Mathematical Programming* and *Journal of the Association of Computing Machinery*, several editorial advisory boards, and reviewing for about 40 journals. Through his public lectures, he has often exhorted the OR/MS community to understand and adapt to changes (cf., [11]). His professional society service includes the presidency of The Institute of Management Sciences (TIMS) in 1981–1982 and of INFORMS in 1997. In 1982 he founded the Management Science Roundtable, an organization composed of the leaders of OR/MS activity in about 50 companies, and he remains actively involved.

Not surprisingly, Professor Geoffrion's research and service has earned him many accolades. He is an honorary member of Omega Rho, a Fellow of the International Academy of Management, a Fellow of INFORMS, and a member of the National Academy of Engineering. In 1992 he was awarded the Distinguished Service Medal from TIMS, in 2000 the George E. Kimball Medal from INFORMS, in 2002 the Harold Larnder Memorial Prize from the Canadian Operational Research Society, and in 2005 an honorary doctorate from RWTH Aachen University (Germany).

References

1. Dolk D (2010) Structured modeling and model management. In Sodhi M, Tang CS (eds) A long view of OR/MS research and practice. Springer, New York
2. Geoffrion AM (1970) Elements of large-scale mathematical programming. Part I: Concepts. Management Science 16(11):652–675

3. Geoffrion AM (1971) Duality in nonlinear programming: A simplified applications-oriented development. SIAM Review 13(1):1–37
4. Geoffrion AM (1974) Lagrangian relaxation for Integer programming. Mathematical Programming Study 2:82–114
5. Geoffrion AM (1976) The purpose of mathematical programming is insights, not numbers. Interfaces 7(1):81–92
6. Geoffrion AM (1987) An introduction to structured modeling. Management Science 33(5):547–588
7. Geoffrion AM (1989) The formal aspects of structured modeling. Operations Research 37(1):30–51
8. Geoffrion AM (1991) FW/SM: A prototype structured modeling environment. Management Science 37(12):1513–1538
9. Geoffrion AM (1992a) The SML language for structured modeling: Levels 1 and 2. Operations Research 40(1):38–57
10. Geoffrion AM (1992b) The SML language for structured modeling: Levels 3 and 4. Operations Research 40(1):58–75
11. Geoffrion AM (1992c) Forces, trends, and opportunities in MS/OR. Operations Research 40(3):423–445
12. Geoffrion AM, Graves G (1974) Multicommodity distribution system design by Benders decomposition. Management Science 20(5):822–844
13. Geoffrion AM, Krishnan R (2001) Prospects for operations research in the e-business era. Interfaces 31(2):6–36
14. Geoffrion AM, Krishnan R (2003a) E-business and management science: Mutual impacts (part 1 of 2). Special issue on e-business and management science. Management Science 49(10):1275–1286
15. Geoffrion AM, Krishnan R (2003b) E-business and management science: Mutual impacts (part 2 of 2). Special issue on e-business and management science. Management Science 49(11):1445–1456
16. Geoffrion AM, Marsten R (1972) Integer programming algorithm: A framework and state-of-the-art survey. Management Science 18(9):465–491
17. Geoffrion AM, Powers R (1995) Twenty years of strategic distribution system design: An evolutionary perspective. Interfaces 25(5):105–127
18. Powers RF, Karrenbauer JJ, Doolittle G (1983) The myth of the simple model. CPMS/TIMS Prize Papers. Interfaces 13(6):84–91

Contributors

Mustafa Atlihan
LINDO Systems, 1415 N. Dayton Street, Chicago, IL 60622, USA

Manel Baucells
Department of Managerial Decision Sciences, IESE Business School, Barcelona, Spain

Dirk Beyer
M-Factor, Inc., San Mateo, CA 94404, USA

John R. Birge
Booth School of Business, University of Chicago, Chicago, IL, USA

Gerald G. Brown
Department of Operations Research, Naval Postgraduate School, Monterey, CA 93943, USA

Kay-Yut Chen
HP Labs, Palo Alto, CA, USA

Scott Clearwater
HP Labs, Palo Alto, CA, USA

Kevin Cunningham
LINDO Systems, 1415 N. Dayton Street, Chicago, IL 60622, USA

Mark S. Daskin
University of Michigan, USA

Daniel Dolk
Department of Information Sciences, Naval Postgraduate School, Monterey, CA 93943, USA

Donald Erlenkotter
Anderson Graduate School of Management, University of California, Los Angeles, CA, USA

Qi Feng
McCombs School of Business, University of Texas at Austin, Austin, TX, USA

Robert Fourer
Northwestern University, Evanston, IL, USA

A. M. Geoffrion
University of California, Los Angeles, CA, USA

G. W. Graves
University of California, Los Angeles, CA, USA

Bernardo A. Huberman
HP Labs, Palo Alto, CA, USA

Shailendra Jain
HP Labs, Palo Alto, CA, USA

Zainab Jamal
HP Labs, Palo Alto, CA, USA

Gautier Laude
LINDO Systems, 1415 N. Dayton Street, Chicago, IL 60622, USA

Hau L. Lee
Graduate School of Business, Stanford University, Stanford, CA 94305, USA

Grace Lin
World Resource Optimization Inc., Chappaqua, NY, USA; IBM Global Business Services, Armonk, NY, USA

Sanjay Mehrotra
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA

Richard Powers
Formerly at Insights, Inc., Stuart, FL, USA

Richard E. Rosenthal
Department of Operations Research, Naval Postgraduate School, Monterey, CA 93943, USA

Rakesh K. Sarin
Decisions, Operations & Technology Management Area, UCLA Anderson School of Management, University of California, Los Angeles, Los Angeles, CA, USA

Linus Schrage
University of Chicago, Chicago, IL, USA

Alper Sen
Department of Industrial Engineering, Bilkent University, Ankara, Turkey

ManMohan S. Sodhi
Cass Business School, City University of London, 106 Bunhill Row, London EC1Y 8TZ, UK

Christopher S. Tang
UCLA Anderson School, UCLA, 110 Westwood Plaza, Los Angeles, CA 90095, USA

Hsiu-Khuern Tang
Intuit, Moutain View, CA, USA

Bob Tarjan
HP Labs, Palo Alto, CA, USA

Jonathan Turner
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA

Krishna Venkatraman
Intuit, Moutain View, CA, USA

Ko-Yang Wang
World Resource Optimization Inc., Chappaqua, NY, USA; IBM Global Business Services, Armonk, NY, USA

Julie Ward
HP Labs, Palo Alto, CA, USA

Alex Zhang
HP Labs, Palo Alto, CA, USA

Bin Zhang
HP Labs, Palo Alto, CA, USA

Chapter 1
Introduction: A Long View of Research and Practice in Operations Research and Management Science

ManMohan S. Sodhi, Christopher S. Tang

1.1 The Roots of Operations Research

Operations Research (O.R.) is rooted in three fields: military operations, economics, and computer science. Operations Research (O.R.)—or, Operational Research—as a field was formally created by scientists in the UK, in particular by researchers working for the Royal Air Force. At the same time, there were parallel efforts in the US to examine ways of making better decisions in the different areas of military operations during WWII [15]. Still, research in operations already had a long history in England rooted in economics, going back to Charles Babbage's study of the pin industry (that following Adam Smith's "division of labor" study of the same industry) and of the postal system resulting in "penny post" that continues to be the model in most countries, thus justifiably earning Babbage the "father of operational research" [23]. It is interesting that Babbage also designed the analytic engine, essentially a programmable computer, because modern O.R.'s insistence on mathematical theory lie in the work of von Neumann and Alan Turing among others who laid down the foundations of the modern computer and of computer science. This book, with a long view of research and practice in O.R., reflects these three roots of operations research.

We can view O.R. as a kind of "management engineering"; in fact the name "management science" co-evolved and the field is sometimes called "operations research/management science" (OR/MS). In this, it follows the path of many engineering fields having originated as military engineering over the past two centuries. The success of OR/MS military applications motivated others to develop and apply OR/MS tools to solve similar problems arising in industry starting in the late 1940s. Many companies created OR/MS departments for internal consulting. Gradually, many engineering and business schools created new groups and

ManMohan S. Sodhi
Cass Business School, City University of London, 106 Bunhill Row, London, EC1Y 8TZ, UK

Christopher S. Tang
UCLA Anderson School, UCLA, 110 Westwood Plaza, Los Angeles, CA 90095, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_1,

programs—OR, MS, Operations Management, Decision Sciences, System Engineering, etc.—to meet the need for OR-trained graduates and better OR methods. OR/MS continued to flourish during 1970s and 1980s in universities and in industry despite questions about the directions of development within the community [16, 17].

Since the 1950s, OR/MS has expanded rapidly both in terms of the application domains and in terms of modeling and solution approaches, drawing strengths from its three roots. Growing from a group of researchers solving military problems, the field now has a well-developed community comprising of practitioners and academics developing modeling approaches and tools for solving problems arising in different functional areas, e.g., finance, marketing, and operations, and in different sectors, e.g., manufacturing, telecommunications, and government. The domains of OR/MS applications rooted in military logistics alone expanded to production planning, distribution planning, and eventually to global supply chain planning. Likewise, the focus on manufacturing or transportation operations broadened to include health care, finance, and many other fields. At the same time, the underlying modeling and solution approaches have evolved from deterministic to stochastic models [5]. The computing platforms also diversified, starting from the mainframe to minicomputers, personal computers, or even mobile computing platform [13]. Finally, on the economics front, the objectives for improvement have evolved from simple single-firm-single-objective to multi-firm-multi-objective models and a typical journal article will encompass the divergent objectives of multiple players.

1.2 About This Compilation

This book is divided into two sections, the first section with chapters taking a long view of the *past* few decades and the second section with chapters taking a long view of the *future*. The first section sheds light on where we are and how we got here and the second section provides opportunities for application and research for the coming decades. Our concluding chapter attempts to span both, viewing the community of OR professionals—practitioners, researchers and teachers—as an ecosystem in which the evolution of OR has taken place and can continue to thrive to take advantage of these opportunities.

1.3 Part I—A Long View of the Past

The chapters in Part I take a retrospective look spanning decades.

1.3.1 Use of OR for Economic Development

Use of OR for economic development goes back quite far although Leontief [19, 20] devised “input–output” modeling. As a result many countries adopted input–output

modeling. Over time, this also gave impetus to application of a broader base of OR tools for economic development. Consider, for example, India. Erlenkotter [8] provides an account of modeling applications from the 1960s, considered "large-scale" in those times, to explore options for the economic development of India. Erlenkotter's account includes institutional environment, application and evolution of the models, and political and economic ramifications thus capturing a reality of OR/MS that is rare in the professional literature describing such models and their application.

1.3.2 The Principal Approaches for Solving Large-Scale Mathematical Programs

As computers become more powerful and efficient, OR professionals (practitioners and researchers) are aspired to solve real-world problems that can involve millions of decision variables. Consequently, there is a constant need to develop more efficient approaches for solving large-scale mathematical programming problems. Birge [4] provides a thoughtful review of fundamental methods for solving large-scale problems that are based on three principal approaches described in Geoffrion [10], namely, *projection*, *outer linearization*, and *inner linearization*. In addition, Birge establishes a link between these three approaches and recent advances in mathematical programming and how they form a basis for solving a variety of real-world problems.

1.3.3 Efficient Distribution System Designs

Distribution system design typically involves the optimal location of intermediate distribution facilities between plants and customers. Geoffrion and Graves [12], whose paper is reprinted here, presented a multi-commodity capacitated single-period version of this problem as a mixed integer linear program. They developed a solution technique based on Benders Decomposition and describe its implementation and application for a major food manufacturing company and obtained a provably optimal solution with a surprisingly small number of Benders cuts. Their method provided a computationally efficient technique that became the basis of application of math programming models to large-scale problems in industry and government; see for instance, Geoffrion and Powers [14] who described the subsequent evolution of distribution design system over the period between 1976 and 1995.

1.3.4 Modeling and Modeling Frameworks

Dolk [7] offers a historical perspective on modeling and model management systems. He uses Geoffrion's Structured Modeling [11], developed in the 1980s, to

address such questions as, Is model management relevant? Can we reframe the basic objectives of such research in today's network-driven, simulation-centric technologies? Answers to these questions remain relevant today in guiding further development of modeling systems.

1.3.5 Distribution and Supply Chain Planning from 1985 to 2010

Next, Powers [22] shares his perspectives regarding the evolution OR/MS applications to logistics planning systems from 1985 to 2010, 25 years of applying OR/MS to corporations and governments all over the world. He argues that the impact of this work resulted in top companies recognizing the value of OR/MS in making resource allocation decisions.

1.3.6 Insight from Application

Providing decision support in the real world is difficult because it necessarily requires dealing with enterprise data systems, legacy procedures, and people with agendas different from the one you are charged with. Brown and Rosenthal [6] provide key insights obtained from his field experience of completing hundreds of optimization-based decision-support engagements over several decades.

1.4 Part II—A Long View of the Future

The other contributing authors present emerging trends for future development of OR/MS tools and applications.

1.4.1 Extending Modeling Interfaces to Deal with Uncertainty

Increasing perception of risk and improved computation technology have resulted in extension of mathematical programming models to stochastic programming. However, tools for modeling practical situations using stochastic programming and thereby creating a broad base of experience are still in short supply. Atlihan et al. [1] describe the stochastic programming (SP) capabilities added to LINDO API (Application Programming Interface) optimization library, as well as how these SP capabilities are presented to users in the modeling systems What'sBest! and LINGO. They discuss the features needed to make SP both easy to use and yet powerful. For instance, they discuss generality in terms of number of stages of the stochastic programming model and allowing integer variables in any stage. Constraints may be linear or nonlinear. Achieving such goals is a challenge because of adding stochastic features to already difficult deterministic optimization problems. They discuss

how developers of such systems need to decide where a particular computational capability should reside: in the frontend that is seen by the user or in the computational engine that does the "heavy computational lifting."

1.4.2 Extending Applications in the Supply Chain

This chapter presents four applied research projects that extend supply chain applications [3]. These projects are being undertaken by the Business Optimization Lab of Hewlett-Packard (HP) Labs to address HP's business needs in diverse areas. The first project describes HP Labs' work in *product variety management*, which is at the interface of marketing and supply chain management decisions. HP Labs introduced a new metric, coverage, for evaluating product portfolios in configurable product businesses and an accompanying Revenue Coverage Optimization tool (RCO). The project focuses on developing *prediction markets for forecasting business events*, involving a handful of busy experts, who do not constitute an efficient market. The work entails harnessing the distributed knowledge of these experts using a two-stage mechanism. The third project encompasses *modeling of rare events for the purpose of marketing*, for instance, to estimate the response probabilities at the customer level to a direct mail campaign when the campaign sizes are very large (in millions) and the response rates are extremely low. The fourth project involves a mathematical programming model that is the core of a number of *decision-support applications* that range from design of manufacturing and distribution networks to evaluation of complex supplier offers in logistics procurement processes.

1.4.3 Global Trade

To sustain profitable growth, many multinational firms focused on two basic strategies. To reduce cost, many firms source from developing countries. To increase revenue, these firms are also selling in various development countries because of their market potentials. To operate these global supply chains effectively, one needs to align the operations of these supply chains with the global trade process. Lee [18] describes how trade agreements, regulations, and local requirements can affect supply chain efficiency. Also, he explains how process re-engineering and information technologies can be helpful in reducing the logistics frictions involved in the global trade processes.

1.4.4 Globally Integrated Enterprises

As more multinational firms launch their global initiatives, many firms find it difficult to obtain competitive advantages mainly due to "the world is flat" syndrome.

To compete successfully in the global marketplace, firms need to differentiate themselves by creating unique value. To do so, Lin and Wang [21] argue that multinational firms must make structural, operational, and cultural changes. Using IBM as a case in point, they show how IBM has transformed itself from a high-tech firm to a "globally integrated enterprise" that utilizes global resources to compete globally without losing sight on its social and environmental responsibilities.

1.4.5 The Internet

Fourer [9] describes three types of projects that fall into the intersection of cyberinfrastructure and large-scale optimization. First, there are the frameworks for making optimization software more readily available. Second, there are projects related by the goal of helping people make better use of available optimization software. Finally, there are projects that apply diverse high-performance computing facilities to problems of optimization. He presents these as having an encouraging future, especially in the context of emerging business models.

1.4.6 Health Care

With ageing population in the developed countries and "western-style" diseases on the rise in emerging economies, health care is an area of national importance in countries around the globe. Turner et al. [26] review resource management as an important area within health care because of the system's unique objectives and challenges. They review recent papers in planning and scheduling along four dimensions: (a) who or what is being scheduled, (b) the planning or scheduling horizon, (c) the level of uncertainty inherent in the planning, and (d) the decision criteria. They point out that the problems at the extreme ends of the planning/scheduling horizon deserve more attention: long-term planning/staffing and real-time task assignment.

1.4.7 Happiness

As societies around the world are getting more affluent, questions are increasingly arising about the pursuit of happiness. Studies have suggested that happiness or even "satisfaction" remained flat over the past few decades (reference? Economist?) even as personal wealth or income has risen, thus raising questions about "utility" as a monotonically increasing function of wealth. Baucells and Sarin [2] seek to explain this anomaly and key empirical findings in the happiness literature. They consider a resource allocation problem in which time is the principal resource. Utility is derived from time-consuming leisure activities, as well as from consumption that comes from time-consuming income-generating activities. They examine the impact

of projection bias on time allocation between work and leisure and show how this bias can cause an individual to overrate the utility derived from income, causing him to allocate more than the optimal time to work and producing a scenario in which a higher wage rate results in a lower total utility.

1.4.8 The OR/MS Ecosystem as the Context for the Future

Based on the collected thoughts of many researchers, we wrap up this book with our perspectives about the future of OR/MS as an ecosystem [25] based on an earlier paper [24]. While research and practice in OR/MS is flourishing, we believe that as a whole the area is at threat in that research, teaching, and practice are becoming increasingly disengaged from each other in the OR/MS ecosystem. This ecosystem comprises researchers, educators, and practitioners in its core along with end users, universities, and funding agencies. It is possible that OR/MS in the future will occupy only niche areas but disappear as a distinct field even though its tools would live on. We present the ecosystem's strengths, weaknesses, opportunities, and threats before discussing the activities the community needs to undertake to mitigate threats and overcome weaknesses so as to use our strengths to exploit the opportunities that lie ahead. These activities can strengthen the interactions among different interest groups of our OR/MS ecosystem, creating a virtuous cycle associated with healthy flows between the various communities in the OR/MS ecosystem.

References

1. Atlihan M, Cunningham K, Laude G, Schrage L (2010) Challenges in adding a stochastic programming/scenario planning capability to a general purpose optimization modeling system. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 117–135
2. Baucells M, Sarin R (2010) Optimizing happiness. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 249–273
3. Beyer D, Clearwater S, Chen KY, Feng Q, Huberman BA, Jain S, Jamal A, Sen A, Tang HK, Tarjan B, Venkatraman K, Ward J, Zhang A, Zhang B (2010) Advances in business analytics at HP laboratories. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 137–173
4. Birge J (2010) The persistence and effectiveness of large-scale mathematical programming strategies: Projection, outer linearization, and inner linearization. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 23–33
5. Birge J, Louveaux F (1997) Introduction to stochastic programming. Springer, New York, NY
6. Brown G, Rosenthal RE (2008) Optimization tradecraft: Hard-won insights from real-world decision support. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer New York, NY, pp. 99–114. (Reprinted with permission from INFORMS)

7. Dolk D (2010) Structured modeling and model management. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 63–88
8. Erlenkotter D (2010) Economic planning models for India in the 1960s. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 11–22
9. Fourer R (2010) Cyberinfrastructure and optimization. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 219–229
10. Geoffrion AM (1970) Elements of large-scale mathematical programming, Part I: Concepts. Management Science 16(11):652–675
11. Geoffrion AM (1987) An introduction to structured modeling. Management Science 33(5):547–588
12. Geoffrion AM, Graves G (1974) Multi-commodity distribution system design by Bender's decomposition. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY. (Reprinted with permission from INFORMS)
13. Geoffrion AM, Krishnan R (2003) E-business and management science: Mutual impacts (Part 1 of 2). Special issue on e-business and management science. Management Science 49(10):1275–1286
14. Geoffrion AM, Powers R (1995) Twenty years of strategic distribution system design: An evolutionary perspective. Interfaces 25(5):105–127
15. Kirby MW (2000) Operations research trajectories: The Anglo-American experience from the 1940s to the 1990s. Operations Research 48(5):661–670
16. Kirby MW, Capey R (1998) The origins and diffusion of operational research in the UK. Journal Operational Research Society 49(4):307–326
17. Kirkwood CW (1990) Does operations research address strategy? Operations Research 38(5):747–751
18. Lee HL (2010) Global trade process and supply chain management. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 175–193
19. Leontief W (1936) Quantitative input and output relations in the economic system of the United States. Rev Economics Statistics 18(3):105–125
20. Leontief W (1966) Input-output economics. Oxford University Press, New York, NY
21. Lin G, Wang KY (2010) Sustainable globally integrated enterprises. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 195–217
22. Powers R (2010) Retrospective: 25 years of applying management science to logistics. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 89–98
23. Sodhi M (2007) What about the "O" in O.R.? OR/MS Today (December). Retrieved from http://www.lionhrtpub.com/orms/orms-12-07/frqed.html on 8th Feb 2010
24. Sodhi M, Tang CS (2008) The OR ecosystem: Strengths, weaknesses, opportunities and threats. Operations Research 56(2):267–277
25. Sodhi M, Tang CS (2010) Capitalizing on our strengths to avail opportunities in the face of weakness and threats. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 275–297
26. Turner J, Mehrotra S, Daskin MS (2010) Perspectives on healthcare resource management problems. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science: The past and the future. Springer, New York, NY, pp. 231–247

Part I
A Long View of the Past

Chapter 2
Economic Planning Models for India in the 1960s

Donald Erlenkotter

Abstract In the 1960s two major linear programming models were constructed to provide guidance for planning the economic development of India. These multisectoral, multiperiod models, although modest in size compared to present linear programming applications, were regarded as large according to the standards and computing capabilities of that time. We review the experiences with these two applications and discuss how they demonstrate the need for Geoffrion's subsequent research in large-scale mathematical programming, data aggregation in models, and structured modeling.

2.1 Preface

The early and seminal work in mathematical programming by Art Geoffrion included major contributions in three important areas: large-scale programming, data aggregation in models, and structured modeling. Through large-scale programming approaches, specific model structures are exploited to enable solution of much larger problems than would be possible with standard methods. Data aggregation seeks to reduce model size by justifiable combination of activities and their data into aggregate activities. Structured modeling stresses the separation of the actual mathematical model from its specific realization in data.

Here we provide an account of some modeling applications from the 1960s that were considered as large scale by the standards of the time. This experience provides insight into the need for innovations of the sort subsequently developed by Geoffrion. These models were designed to explore options for the economic development of India, a country then with some 500 million people. Our account covers the total modeling experience as it evolved, including institutional environment, application and evolution of the models, and political and economic ramifications. Real applications of models invariably are linked to such broader

Donald Erlenkotter
Anderson Graduate School of Management, University of California, Los Angeles, CA, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_2,

contexts, even though this is often excluded from the professional literature describing the models.

2.2 Introduction

In 1966 I went to India to work on sectoral and industrial planning studies for the US Agency for International Development (USAID) Mission in New Delhi. This work was in support of projects that USAID had under consideration for financing. During my 3-year assignment in India, I became involved in the national economic modeling effort that was conducted to explore the potential impact of different levels of economic assistance on India's development. Here I discuss the use and evolution of these national economic models in India from the perspective of my experiences. Most of what I say about modeling efforts there prior to 1966 is based on recollections of contemporary conversations and experiences with those who were close to these efforts and who had no reason to give biased views. These recollections correspond reasonably well with published accounts of this work.[1]

The use of national planning models in India for exploring growth options had its heritage in the simple growth models developed by Frank Ramsey in the 1920s.[2] This type of model is solved by the calculus of variations. While such models provide some insight into the relationship between savings and growth, they are far from adequate as guides to economic policy. Growth models are heavily dependent on one magical parameter, the capital–output ratio. In reality, there are distinct capital–output ratios for each economic sector and the allocation of investment among these sectors influences the overall capital–output ratio. Allocation of investment among sectors also implies decisions about imports and exports and so one is led to expand models to include international trade possibilities.

Multisectoral economic growth models became feasible with the development of computer codes for solving mathematical programming problems. The first models of this sort were devised by Ragnar Frisch of Norway and Jan Tinbergen of The Netherlands in the 1950s, and in 1969 these two men shared the first Nobel Memorial Prize in Economic Sciences for their work. The underlying structure of these models was based on the interindustry input–output framework devised at Harvard by Wassily Leontief, for which he received the Nobel Memorial Prize in Economic Sciences in 1973.

2.3 The MIT Model for India

In the early 1960s, a project was launched to develop and apply such models in India. At the time, India had carried out, more or less, a series of 5-year plans beginning from 1951 and was the largest experiment in economic planning in the non-totalitarian world. In reality, these plans were far removed from the rigid format of their counterparts in the Soviet Union, and I don't believe that anyone expected an

economic planning model to provide an exact prescription for action. These models were intended more as information systems that would provide guidance as to the potential impact of various policy options.[3]

The initial modeling effort in India was launched by the Massachusetts Institute of Technology's Center for International Studies, which was located in Cambridge, Massachusetts, with a branch office in New Delhi. The project team was international, with leadership provided by Sukhamoy Chakravarty, Richard Eckaus, Louis Lefeber, and Kirit Parikh.[4] For short, their model was known as the CELP model. India then had little in the way of resources for high-speed computing and so the project team was divided into two groups. Chakravarty and Lefeber were mainly in India and they had the primary responsibility for data acquisition. Eckaus and Parikh were in Cambridge and they were in charge of carrying out the computations. International communications were not easy at this time, since mail was slow and telephone service was erratic and very expensive. Communications difficulties were to play a critical role in the outcome of this modeling exercise.

In any modeling effort, the model is regarded as "on probation" until its structure and data have been thoroughly checked and the model's results are understood and regarded as reliable. As data were acquired for the CELP model, preliminary runs were being made at MIT. In October 1964, during these runs of the model and while the data were still being checked, the MIT Center in Cambridge was visited by India's Ambassador to the United States, B. K. Nehru.[5] The ambassador was very interested in the model and its results, and when he returned to Washington, he sent a cable back to New Delhi reporting his findings. Then the fun began.

At that time, India was in the process of formulating its Fourth 5-Year Plan, which was intended to span the period from 1966 to 1971. There were two major factions involved in the preparations for this plan. The Planning Commission was responsible for the final dimensions of the plan. In particular, the detailed parameters underlying the plan were overseen by the Perspective Planning Division, headed by Pitambar Pant. The Planning Commission generally favored an "ambitious" plan with high-growth targets, since the need for rapid development in India was obvious. The other major faction was represented by the Ministry of Finance (MoF), which had the responsibility for raising the resources necessary to carry out the plan. Not surprisingly, the MoF tended to favor a less ambitious plan than did the Planning Commission.

The Indian Ambassador in Washington was aligned with the MoF faction. He had reported back to New Delhi that the MIT experts' calculations showed the Planning Commission's announced targets for the Fourth Plan could not be attained. This, of course, provided major support for the MoF's campaign for a less ambitious plan. And, not surprisingly, these latest developments in the ongoing controversy over the plan soon appeared in the press.

On the other hand, Chakravarty and Lefeber, in New Delhi, had been working closely with the Planning Commission and they immediately lined up on that side of the dispute. The computer runs at MIT, they said, were preliminary and hadn't used the most recent data available in India. In particular, there was one crucial

and difficult-to-estimate parameter that made a significant difference in the model's results. This was the capital–output ratio for the housing sector, which is a substantial portion of the Indian economy. Output for housing typically is an imputed figure, and a number of assumptions must be made to arrive at an imputation. Once the data were adjusted, the Planning Commission's targets actually were reasonable, in the opinion of Chakravarty and Lefeber.

The impact of press involvement on the modeling process was devastating. In a reaction typical for India, the next charge was that the MIT Center was a front for CIA espionage in the country and that a large safe in the Center's New Delhi office was used to store clandestine intelligence materials. As this political storm grew, the New Delhi office was closed and the Center's operations in India ceased. The project team split, with Eckaus and Parikh publishing a book on their modeling efforts[6] while Chakravarty and Lefeber published separately in India on their work.[7] According to Rosen, the alleged CIA involvement here "helped to start a process leading to a more or less steady decline of opportunities for academic social science (and economic) research by American scholars in India."[8]

2.4 The Manne–Weisskopf Model for India

I arrived in New Delhi in August 1966 from Stanford University, where I had been working on my Ph.D. dissertation. Already there was Alan Manne, my dissertation advisor at Stanford, who had come on a 1-year assignment with USAID as the economic adviser to the Mission Director, John P. Lewis. Alan had been in India 2 years before with the MIT Center, working on sectoral planning studies involving the sizes, locations, and time phasing of plants in various industries. He and I would continue that line of work. In addition, following another research track initiated during his earlier stay in India, he would establish a multiperiod, multisectoral national planning model that could be used to explore the impact of different economic assistance strategies.[9]

Scheduled to join us was Thomas E. Weisskopf, who had been finishing his dissertation at MIT on a programming model for import substitution for India.[10] However, by the time I had reported to Washington for my USAID orientation, Tom had resigned his position in protest over the US bombing of Hanoi and Haiphong. Under a last-minute arrangement, he came over to join the Planning Unit of the Indian Statistical Institute in New Delhi. There he would carry out economic modeling work as one of his assignments.

Alan and Tom began work on the dynamic multisectoral (DMS) model for India in close association with Pitambar Pant and the Perspective Planning Division, which was a primary source of data. India's Fourth 5-Year Plan had been delayed for several years due to the 1965 war with Pakistan and two successive years of disastrous droughts. The model would span the time interval from 1967 through 1975, which included the revised Plan period. It differed from previous efforts both in its scope and in the incorporation of new theoretical ideas that Manne had developed to enable a model with just a few periods to approximate reasonably well the reality of

an unlimited horizon.[11] It also employed the objective of maximizing a "gradualist" consumption path, which increased at an increasing rate over time. Although I was involved mainly with industry studies at the time, I kept abreast of the work on the DMS model.

The DMS model was not large by present-day standards, but in 1966 it required what was considered a very large computer.[12] There was no such computer in New Delhi at the time. The Ford Foundation had brought in several IBM 1620s, but these were much too small for our purposes. The Indian Institute of Technology at Kanpur had an IBM 7044, but this was an inconvenient site and the availability of software was problematical. Our choice for a computational facility was the Tata Institute for Fundamental Research (TIFR) on Colaba Point in Bombay (now Mumbai). TIFR had a Control Data Corporation (CDC) 3600 with a linear programming package known as CDM4.

The initial trial runs with the DMS model began in August 1967 and continued into the following month.[13] On our trips to Bombay, we had to use the 3600 late at night and in the wee hours of the morning since the Institute's physicists had priority. The first runs there were an education for all. In India there was quite a rigid division of tasks among project personnel. The scientists would design the program and hand it to the programmers for coding. The programmers would then give the code to clerical staff for keypunching and if corrections had to be made these would go back to the keypunching staff.[14] This time-consuming process would not work with the limited time we had on our trips to Bombay, especially since there was no clerical staff available at night. The breakthrough came when Alan Manne sat down at the keypunch and banged out cards with the corrections he needed. Our assistants quickly got the message that the work was to be done expeditiously, regardless of who had to do it.

As we began our computer work, I was able to learn some useful information about the CDM4 linear programming code. "CD" was short for "Control Data," obviously. The actual name of the code was "M4." In the summer of 1964 I had worked in the Mathematical Modeling Group at Standard Oil of California (SOCal) in San Francisco. They had a linear programming code named "M3," which was the third generation of codes originally developed jointly by SOCal and the RAND Corporation in Santa Monica. At that time, development of a fairly reliable linear programming code had been at least a million dollar undertaking. The names of various subroutines in the CDM4 code verified its pedigree to me and provided a great deal of information about how the code operated. This turned out to be quite useful later on.

TIFR provided a very pleasant working environment, even late at night. One could look out over the Arabian Sea or watch the rain squalls sweep in. The institute was in a striking modern building, with walls well decorated with contemporary Indian paintings. These were much appreciated as we waited for the whirring tape drives to complete our runs with the model. Even on this large and high-speed computer, and under the best of circumstances, each run could easily take 45 min.

We were able to complete our runs and the results were recorded in a preliminary paper.[15] Manne returned to Stanford shortly after these preliminary runs were made, taking a copy of the model to run there. A paper describing a revised version

of the DMS model with new computational results was presented by Manne and Weisskopf in January 1968 at the Fourth International Conference on Input–Output Techniques held in Geneva. The final version of this paper was published in the proceedings volume for that conference.[16]

In mid-1968 we ran the DMS model again to obtain updated calculations for the impact of various aid levels on the Indian economy. These results were used in the *Country Field Submission* forwarded by USAID-New Delhi to Washington to support the annual aid request.[17] Later that summer Weisskopf left India to join the economics faculty at Harvard University. I continued to maintain and use the model at USAID through mid-1969.

Early in 1969 we had the opportunity to prepare a report for the World Bank (Pearson) Commission on International Development using the DMS model. This report assessed the impact of one billion dollars in additional foreign aid provided over a 10-year period and coupled with a set of socially oriented government expenditure programs intended to attack the "low end" poverty problem in India.[18] For this exercise, the model was modified slightly to incorporate a constant per capita marginal savings rate and a maximand of terminal year net national product.[19] We also extended the model's time frame to include a fifth time period. Although this had seemed to be a relatively straightforward undertaking, the first time we tried to find a solution with the expanded model the computer run exceeded the time available and we had to stop prematurely. This was very puzzling and a postmortem examination revealed that the program had essentially reached a final solution but had failed to terminate normally because of difficulties with minor numerical errors in the computations.

Solving models of this size at TIFR was never a straightforward undertaking, not so much because of inadequacies of the software but more because of hardware breakdowns. Replacement parts often were impossible to obtain because of India's strict controls over imports. Normally we could stop and restart a computational run by saving an intermediate solution. However, this required a free tape drive. Usually there was one free drive in addition to the one needed for saving the intermediate solution, but at the time this spare drive was out of order and would be so until a CDC technician was able to come to India and smuggle in the needed parts. The night we made our run, the operator had sheared off the mounting spindle on one of the other drives while mounting a tape. Without this drive, we could not stop a run and restart—any stop meant starting again from scratch. This is why we had watched the tape drives whir back and forth for several hours without stopping to check the intermediate results.

Analyzing the run log revealed that the program had continued for well over an hour to exchange one variable for another and then to reverse this exchange over and over. This is called "cycling," and any mathematician will provide proof that it is impossible. But mathematicians do not consider numerical error in their analysis, and with a large model there can be enough numerical errors to cause such cycling. Meanwhile, our team at USAID was waiting for our results so the report could be completed in time for the Pearson Commission's deadline.

This is where my knowledge of the CDM4 code saved us. I knew that the code used what is called the "product form of the inverse" to calculate successive trial

solutions for the model. And, the earlier versions of the code I had used at SOCal had naively introduced each variable into the calculations in the order provided in the input. We had placed all the data for the investment variables at the beginning of the input card deck, followed by sets of variables that appeared only in each individual time period. The investment variables linked all the periods together and by including these linking variables first we were causing the variables for the different time periods to be linked together as they were brought in. By simply moving the cards for the investment variables to the end of the deck, we kept the individual period variables disconnected until all the periods had been processed. It turned out that this not only eliminated the problem caused by numerical error, it also substantially reduced the time for each run.[20]

By the time all this was figured out, I had to return to Delhi to wrap up the report. I left my assistant, S. M. Luthra, to complete the runs, keeping my fingers crossed, and flew back. Luthra worked through several nights and returned with all the results and we were able to complete the report on time. For his efforts, he was given an award by USAID. By this time our enchantment with TIFR and Colaba had diminished considerably, especially when we found that taxis couldn't be obtained out there in the very wee hours of the morning and the trek of several miles back to Bombay on foot in the dark was something less than enjoyable.

One of the features of our Pearson Commission report was an exploration of the consequences of imperfect forecasting of aid amounts. In particular, the report demonstrated the value of reduction in uncertainty through long-term commitment of aid levels.[21] As events turned out, uncertainty rather than long-term commitment was to rule the future of the USA's aid to India.

The new variant of the DMS model was used once more in mid-1969 to support the aid requests in that year's *Country Field Submission*. For these calculations, further revisions were made in the model to incorporate information about India's economic performance in 1969/1970 and to introduce recent estimates of underutilized capacity in several industries. Including this underutilized capacity, which was a consequence of the recession induced by the drought years, led to increased short-term productivity for aid.

2.5 Epilogue

Use of the DMS model at USAID did not survive very long after my departure from India in July 1969 and the changes brought by the Nixon administration. Following Nixon's "tilt" to Pakistan in 1971 during the turmoil that led to the creation of Bangladesh, American aid to India was suspended. The American staff at USAID in New Delhi was cut from 260 in 1968 to just 8 at the beginning of 1974.[22]

In the various computational runs with the DMS model, the long-run target for economic growth typically had been set at 8% per year. Short-term growth rates varied with the level of economic assistance, but were substantially lower. India's actual economic performance during the 1970s and 1980s exhibited growth rates more in the range of 4–5% per year. But from 1990 on, economic growth

accelerated, and over the past 3 years India's annual growth in gross domestic product (GDP) has averaged 8.1%—virtually the same as the long-run target used in the model.[23] This accelerated growth is widely attributed to the removal in 1991 of crippling restrictions on trade and investment that had been imposed for many years by the Government of India through its licensing procedures. Such a "liberalization," or decontrol had been encouraged for many years by USAID, the World Bank, and other international and domestic institutions.

Among those involved in the modeling efforts discussed here, Sukhamoy Chakravarty continued his career as a leading economist in India and internationally. He was a member of India's Planning Commission from 1971 to 1977 and served as Chairman of the Economic Advisory Council of the Prime Minister from 1983 until his death in 1990.[24] Richard Eckaus remained at MIT for many years and continued to work on problems of economic development. He was made Ford International Professor there in 1977 and served as head of the Department of Economics in the late 1980s. Louis Lefeber was denied promotion to full professor by the administration at MIT in 1965, reportedly because he had objected to the MIT Center's use of research on Indian planning for political purposes. He moved to Brandeis University and eventually to York University, where he was the founding director of the Center for Economic Research in Latin America and the Caribbean. Kirit Parikh returned to India, where he continued fundamental work in economic modeling, particularly in agriculture and energy. From 1980 to 1986 he was Program Leader of the Food and Agricultural Program at the International Institute for Applied Systems Analysis (IIASA) in Austria. In 1986 he became Founder-Director of the Indira Gandhi Institute of Development Research in Mumbai and was appointed as a member of India's Planning Commission in 2004. Alan Manne returned to Stanford, where he carried out modeling studies on the Mexican economy and then in the 1970s turned to large-scale energy and environmental modeling research. He continued work in this area up to his death in 2005. Tom Weisskopf went on to become a founder and leader of the Radical Political Economics movement, moving from Harvard to the University of Michigan in 1972.

As part of the evaluation of these past modeling efforts, the data for a version of the DMS model were exhumed and incorporated into a program written in the GAMS modeling language.[25] The effects of the advances in modeling and computation over the intervening years were dramatic. Even on a relatively slow desktop computer, compilation and solution of the "large" model by 1960s standards took no more than a couple of seconds, less than the preparation time for just one of the more than 2000 punched data cards required for the original model.

2.6 Concluding Reflections

As we have seen here, advances in modeling and computation over the past half-century have had an enormous impact on the concept of model size. Much of this, of course, is due to dramatic increases in computational speed and memory capacity.

But what of more model-specific innovations? The recent DMS computations were performed by the CPLEX linear programming system, as integrated with the GAMS modeling language. Although this system does not directly include approaches usually classified under the heading of "large-scale mathematical programming," it does exploit model structure and data sparsity through basis inversion techniques that use LU decomposition. This approach is especially well suited to dynamic planning models, which primarily have a "staircase" data structure.

As for advances in modeling, the algebraic statement of the DMS model remains valid and now can be implemented directly and conveniently through modeling languages such as GAMS. But the original structure of the model was reduced in size for computational purposes by using rather ad hoc aggregation procedures. These aggregations were based on preliminary inspections of the structure of the data. Does this violate the principle of separation of model and data or can it be viewed as an example of skillful modeling practice? We leave it to the reader to ponder this issue.

A major innovation brought by modeling languages is the capability for specifying each data element uniquely and then using the language to perform all the required data calculations. This also provides documentation and transparency that were lacking in modeling efforts of the 1960s, where each data coefficient was calculated separately and punched into an 80-column card. The advantages here for avoiding computational errors and the improved capability for performing revisions to the model are evident.

2.7 Notes

1. In particular, see Rosen, G (1985) Western economists and eastern societies: Agents of change in South Asia, 1950–1970. The Johns Hopkins University Press, Baltimore, MD, pp. 101–146; also Blackmer DLM, (2002) The MIT center for international studies: The Founding Years, 1951–1969. MIT Center for International Studies, Cambridge, MA, pp. 175–201.
2. Ramsey, FP (December 1928) A mathematical theory of saving. Economic Journal 38 (152): 543–559.
3. For an excellent overview of economic analysis and modeling in India's planning efforts, see Bhagwati, JN Chakravarty, S (September 1969) Contributions to Indian economic analysis: A Survey. American Economic Review 59 (4): 1–73, Part 2, Supplement.
4. Chakravarty had received his Ph.D. degree from the Netherlands School of Economics under the supervision of Tinbergen. The conceptual foundations of the modeling effort for India are given in chapters by Chakravarty S, Eckaus S, Lefeber L (1964) In: Rosenstein-Rodan PN (ed) Capital formation and economic development. MIT Press, Cambridge, MA.
5. Rosen, op. cit., pp. 133–134. In the following month, Eckaus presented the model at a National Bureau of Economic Research conference on economic

planning and some of the discussants' comments about the model's policy applicability were quite critical. See Eckaus, RS (1967) Planning in India. In. Millikan MF (ed) National economic planning. Columbia University Press, New York, NY, pp. 305–369. On p. 326 of this paper Eckaus states that

"The numerical solutions remain hypothetical exercises. … In particular, I should like to emphasize that I do not presume to be laying down guidelines for Indian policymakers. The empirical results are intended to be illustrative rather than definitive."

6. Eckaus, RS Parikh, KS (1968) Planning for growth: Multisectoral intertemporal models applied to India. MIT Press, Cambridge, MA. Over a 3-year period, the modeling effort had required 150 h of computer time on an IBM 7094, the largest commercially available computer at the time (p. 15, note 12). The book's dust jacket states that "It is of considerable interest to note that the application of the models to Indian planning produces results strongly suggesting that the Third Five Year Plan and a proposed Fourth Five Year plan were not feasible …"
7. See Chakravarty, S Lefeber, L (February 1965) An optimizing planning model. The Economic Weekly (Bombay) 17 (5–7): 237–252; Srinivasan, TN (February 1965) A critique of the optimising planning model. The Economic Weekly (Bombay) 17 (5–7): 255–264. These papers issue several cautions about the inadequacies of the particular model and its results for policy comparisons. Noticeably absent is any mention of the MIT Center, but on p. 237 Chakravarty and Lefeber comment that "… several misunderstandings about the policy implications of the approach have recently arisen."
8. Rosen, op. cit., p. 143.
9. See Manne, AS Rudra, (February 1965) A consistency model of India's fourth plan. Sankhya: The Indian Journal of Statistics Series B 27: 57–144 Parts 1 & 2 for an earlier planning model done at the MIT Center in New Delhi; also Bergsman, J Manne, AS (November 20, 1965) An almost consistent intertemporal model for India's fourth and fifth plans. The Economic Weekly, (Bombay) 17 (47): 1733–1741, and in Adelman, I Thorbecke E (eds) (1966) The Theory and Design of Economic Development. Johns Hopkins Press, Baltimore, MD, pp. 239–261.
10. Weisskopf, TE (December 1967) A programming model for import substitution in India. Sankhya: The Indian Journal of Statistics Series B 29: 257–306 Parts 3 & 4; "Alternative Patterns of Import Substitution in India," in Chenery, H B et al. (eds) (1971) Studies in development planning. Harvard University Press, Cambridge, MA, pp. 95–121. Weisskopf had begun work on this project in 1964–1965 while on a doctoral research fellowship at the New Delhi branch of the Indian Statistical Institute.
11. See Manne, AS (January 1970) Sufficient conditions for optimality in an infinite horizon development plan. Econometrica 38 (1): 18–38.
12. The DMS model divided the Indian economy into 37 sectors and contained four 2-year time periods. This required 228 constraints and 236 variables in the linear programming formulation. Explicit interindustry detail was included only

for 17 endogenous production-oriented sectors. The remaining 20 sectors were consumption oriented or exogenous and had negligible interindustry deliveries. They were aggregated, with their outputs and inputs related to aggregate consumption, investment, and value added. Applying an analogy from the card game of bridge to this aggregation procedure, Manne commented that "One peek is worth two finesses."

13. The team for the second round of DMS computations in September 1967 consisted of myself and S.M. Luthra from USAID, T. Weisskopf from the ISI Planning Unit, and P.N. Radhakrishnan and H.K. Raina from the Perspective Planning Division.
14. In those days, all input was submitted on 80 column punched computer cards, each of which contained a single coefficient with its row and column identification. The modeling languages of today were not even a distant dream back then.
15. See Manne, AS Weisskopf, TE (December 1967) A dynamic multisectoral planning model for India. Discussion Paper No. 34, Indian Statistical Institute Planning Unit, New Delhi, India.
16. Manne, AS Weisskopf, TE (1970) A dynamic multisectoral model for India, 1967–1975. In: Carter, AP Bródy A (eds) Applications of input-output analysis, Vol. 2. North-Holland, Amsterdam, pp. 70–102. The computations reported in the paper were done at Stanford in May 1968 with DMS5, the fifth version of the DMS model. The paper includes the following caveat:

    "Before applying any of our numerical results to check the internal consistency of India's forthcoming official plan documents, it is essential to isolate those discrepancies that arise from differences in technological norms from those that arise from differences in macroeconomic policy viewpoints. The numerical results from DMS are sensitive, not only to assumptions regarding aid inflows and domestic austerity, but also to estimates of the likely improvements in efficiency of resource use. Regrettably, on this subject of efficiency in implementation, there is only a thin line that separates prudent planning from wishful thinking."
17. Annex C of USAID's *FY 1970 Program Memorandum—India*, titled "Development Planning," makes the following comment on p. C-8:

    "DMS does not purport to quantify two considerations—factors that are partially offsetting, and that would become operative at low levels of aid: First, India's import liberalization program might be substantially curtailed or abandoned. Without this program, a substantial element of allocative efficiency would be lost—a factor over and above the loss of comparative advantage included directly within DMS. Second, without the 'soft option' of aid available, the Government of India might take more vigorous actions to solve the country's development problems. We leave it to the reader to allow for these factors."
18. The attraction of such programs to reduce rural poverty in India continues—recently it has been proposed that up to 25 million people be employed in development projects such as building roads, planting trees, and digging irrigation

canals. See Watson, P (August 25, 2005) India moves toward guaranteed jobs program for rural poor. Los Angeles Times p. A3.

19. In the terminology of planning models, this change converted the DMS model into a "closed loop" format from an "open loop" one. See Manne, AS (1974) Multi-sector models for development planning. Journal of Development Economics 1 (1): 43–69.
20. Subsequent generations of linear programming codes incorporated more sophisticated routines that essentially carried out this reordering of variables automatically.
21. Erlenkotter, D Lubell, H (April 1969) Additional foreign aid for socially oriented government programs. Economic Affairs Division, US Agency for International Development, New Delhi, India.
22. Lelyveld, J (June 25, 1974) A case study in disillusion: U.S. aid effort in India. New York Times p. 6.
23. See "Can India Fly?" The Economist, June 3, 2006, p. 13.
24. Sen, A (1993) Sukhamoy Chakravarty: An appreciation. In: Basu K, Majumdar M, Mitra T (eds) Capital, investment and development: Essays in memory of Sukhamoy Chakravarty. Basil Blackwell Ltd., Oxford, UK, pp. xi–xx.
25. Brooke A, Kendrick D, Meeraus A (1992) GAMS: A user's Guide, release 2.25. The Scientific Press, South San Francisco, CA.

Chapter 3
The Persistence and Effectiveness of Large-Scale Mathematical Programming Strategies: Projection, Outer Linearization, and Inner Linearization

John R. Birge

Abstract Geoffrion [19] gave a framework for efficient solution of large-scale mathematical programming problems based on three principal approaches that he described as *problem manipulations*: projection, outer linearization, and inner linearization. These fundamental methods persist in optimization methodology and underlie many of the innovations and advances since Geoffrion's articulation of their fundamental nature. This chapter reviews the basic principles in these approaches to optimization, their expression in a variety of methods, and the range of their applicability.

3.1 Introduction

Optimization methodology development has been characterized by regular and rapid decreases in solution times. At the same time, problem sizes have also increased dramatically. These efficiency gains derive not just from hardware advances but equally from improvements in the underlying methodology (see, e.g., [10]). While such advances continue to expand the reach and effectiveness of optimization methodology in addressing practical decision problems, much of the fundamental properties in these innovations relate to a set of approaches described in Geoffrion [19].

Geoffrion [19] describes three principal *problem manipulations*: projection, outer linearization, and inner linearization. This chapter describes how these approaches relate to many of the more recent advances in mathematical programming and how they form a basis for the consideration of a variety of problems associated with decision modeling in general. In the following sections, I describe each of the

John R. Birge
Booth School of Business, University of Chicago, Chicago, IL, USA

This chapter is dedicated to Arthur M. Geoffrion for his many contributions to operations research, management science, and mathematical programming. The work was supported by The University of Chicago Booth School of Business.

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_3,

basic manipulations and relate them to more recent uses, their applicability, and effectiveness.

3.2 Projection

The fundamental approach in projection is to manipulate the domain of an optimization problem, generally from a higher to a lower dimension, but also from unbounded to bounded domains. As described in Geoffrion [19], a fundamental problem might be described in two sets of variables x and y [1] as

$$\min_{x\in X, y\in Y} f(x,y) \quad \text{s. t.} \quad g(x,y)\le 0, \tag{3.1}$$

where $X\subset \Re^n$, $Y\subset \Re^p$, $f:\Re^n\times\Re^p\to\Re$, and $g:\Re^n\times\Re^p\to\Re^m$. Geoffrion [19] defines a projection of $X\times Y$ to $X\cap V$ where $V=\{x \mid g(x,y)\le 0 \text{ for some } y\in Y\}$, with objective $v(x)$ such that

$$v(x)=\inf_{y\in Y, g(x,y)\le 0} f(x,y), \tag{3.2}$$

creating the equivalent problem to (3.1) in x alone as

$$\min_{x\in X\cap V} v(x). \tag{3.3}$$

As Geoffrion [19] notes, this manipulation forms the basis of several classical methods in mathematical programming, such as Benders' [5] decomposition and Rosen's [26] partitioning method. He also observes that the fundamental basis of dynamic programming, as in Bellman [4] and Dantzig [11], is to use this form of projection with decisions sequentially determined at each stage. Projection also forms the core of more recent methods that are now the most efficient algorithms for various problem structures. *Interior point* methods, for example, which started from early descriptions by Fiacco and McCormick [17], and which spread broadly in applications following Karmarkar's [20] discovery of their efficient application for linear programs, can be viewed as applications of projection.

3.2.1 Projection in Interior Point Methods

To see how interior point methods fit the general projection in (3.3), consider, as an example, the linear program with $g(x,y)=(Ax+Iy-b,-Ax-Iy+b,-Iy)$, so that $g(x,y)=\{x,y \mid Ax+Iy=b, y\ge 0\}$ and $f(x,y)=c^Tx$. Directly using the projection steps above would, of course, yield

$$\min c^Tx \quad \text{s. t.} \quad Ax\le b. \tag{3.4}$$

[1] The roles of x and y are reversed from Geoffrion [19] to be consistent with later descriptions.

Instead of directly reducing (3.2) to (3.4), interior point methods with projection use two additional projection steps to produce a different iteration. They start with a current iterate (x^k, y^k) and search for $(x, y) = (x^k + s, y^k + t)$ for some search direction (s, t). Relative to the current iterate, problem (3.4) is equivalent to

$$\min c^T (x^k + s) \quad \text{s. t.} \quad As \leq y^k, \tag{3.5}$$

or, for $Y^k = \text{diag}(y^k)$,

$$\min c^T s (+c^T x^k) \quad \text{s. t.} \quad (Y^k)^{-1} As \leq \mathbf{1}, \tag{3.6}$$

where $\mathbf{1}$ is a vector of ones. The first application of projection is to make the region in (3.6) compact by projecting the region in (3.6) into a simplex using the *projective transformation*, $z = s/\left(1 - \frac{1}{m+1} e^T (Y^k)^{-1} As\right)$, which also yields an inverse, $s = z/\left(1 + \frac{1}{m+1} \mathbf{1}^T (Y^k)^{-1} Az\right)$, when the denominator is positive. Now, substituting for s and imposing the feasible constraint yields an equivalent feasible region in z as

$$L^k = \{z \mid A^k z \leq \mathbf{1}\}, \tag{3.7}$$

where

$$A^k = \begin{pmatrix} (Y^k)^{-1} A - \mathbf{1} \frac{1}{m+1} \mathbf{1}^T (Y^k)^{-1} A \\ -\frac{1}{m+1} \mathbf{1}^T (Y^k)^{-1} A \end{pmatrix}.$$

The algorithm then make an appropriate approximation of $c^T s$ (through some form of linearization, the theme discussed in the next section) as $(d^k)^T z$ to yield a surrogate problem

$$\min (d^k)^T z \quad \text{s. t.} \quad A^k z \leq \mathbf{1}, \tag{3.8}$$

which with $w = A^k z$ can in turn be written as

$$\min (\lambda^k)^T w \quad \text{s. t.} \quad w \leq \mathbf{1}, w - \Pi_{A^k} w = 0, \tag{3.9}$$

where $d^k = (A^k)^T \lambda^k$ and Π_{A^k} corresponds to projection onto the column space of A^k. The representation in (3.9)shows how the algorithm can be interpreted as optimizing a linear function over a simplicial region. The algorithm then solves for a search direction over a restriction on (3.9) that can be given as the inner sphere, $W = \{w \mid w^T w \leq m + 1\}$, intersected with the projection constraint. Since a magnification by m of this inner sphere circumscribes the region in (3.9), with a consistent definition of the objective approximation, each step can then be shown to attain a fixed rate of convergence that yields a computationally efficient method overall.

3.2.2 *Projection in Discrete Optimization*

Another significant application area for projection has been in discrete optimization. In this case, the basic problem in (3.1) is assumed to correspond to $Y = \{y_j \mid y_j \in$

$\{0,1\}, j = 1,\ldots,p\}$. The methods first raise the dimension of the problem and then project back into the original dimension to obtain stronger linear approximations of the original feasible region than would be obtained with a direct continuous relaxation of the original problem. The *lift-and-project* method by Balas et al. [2] and similar approaches from Lovász and Schrijver [23] and Sherali and Adams [28] all have this feature.

The lift-and-project method proceeds again with a set of linear constraints $g(x,y) = Ax + Cy - b$, which here includes restrictions implying that $x \geq 0$, $y \geq 0$, and $y \leq \mathbf{1}$. Let $K = \{(x,y) \mid g(x,y) \leq 0\}$ and $K^0 = \{(x,y) \mid (x,y) \in K, y \in Y\}$, i.e., such that $y_j \in \{0,1\}$, for $j = 1,\ldots,p$. The method constructs a sequence of approximations $P_j(K)$ through the following procedure:

1. Let $K' = \{(x,y) \mid (1-y_j)(Ax+Cy-b) \leq 0, y_j(Ax+Cy-b) \leq 0\}$.
2. Replace all $x_i y_j$ terms in the constraints of K' by a new variable u_i, all $y_i y_j$, $i \neq j$ terms by a new variable v_i, and y_j^2 by y_j. Let the resulting feasible region (a polyhedron) in (x,y,u,v) be $M_j(K)$.
3. Project $M_j(K)$ onto (x,y)-space as $P_j(K) = \{(x,y) \mid \exists u,v \text{ s. t. } (x,y,u,v) \in M_j(K)\}$.

By defining these projections iteratively as $P_{1,\ldots,p}(K) = P_1(P_2(\cdots(P_p(K))\cdots))$, it can be shown (Corollary 2.3 in Balas et al. [2]) that $P_{1,\ldots,p}(K) = \text{co}(K^0)$, the convex hull of K^0. By iteratively generating facets of $P_j(K)$, a finite algorithm can then be obtained that follows the basic procedure below (Theorem 3.1 in Balas et al. [2]).

Lift-and-Project Cutting Plane Algorithm

1. Let $K^1 = K$. $k = 1$.
2. Solve for $(x^k, y^k) = \arg\min\{c^T x + d^T Y \mid (x,y) \in K^k\}$. If $y^k \in Y$, stop.
3. Let j be the largest index when $0 < y_j^k < 1$. For $\alpha^k(x,y) \leq b^k$, a facet identified on $P_j(K)$, let $K^{k+1} = K^k \cap \{(x,y) \mid \alpha^k(x,y) \leq b^k\}$.
4. Set $k = k+1$ and go to Step 2.

3.3 Outer Linearization

The lift-and-project cutting plane algorithm in the previous section involves both projection in the construction of the $P_j(K)$ relaxations and also outer linearization through the progressive identification of facets and their inclusion into the kth iterate feasible-region relaxation, K^k. As Geoffrion [19] observes, projection is often combined with outer linearization in the form of the cutting planes as, for example, used in the lift-and-project method.

Broadly, for a problem defined as

$$\min_{x \in X} f(x) \quad \text{s. t.} \quad g(x) \leq 0, \tag{3.10}$$

outer linearization can apply either to the objective (or some part of the objective) or to the constraints. In this way, $f(x)$ is replaced by $f^q(x) = \max_{i=1,\ldots,q}(\alpha_i^T x + \beta_i)$, where $f(x^i) = \alpha_i^T x^i + \beta_i$, and $g(x) \leq 0$ is replaced by $g^r(x) \leq 0$, where $g_i^r(x) = E_i^T x - e_i \leq 0$ for $i = 1,\ldots,r$. For these linearizations to be *outer*, $f^q(x) \leq f(x)$ and $\{x \mid g^r(x) \leq 0\} \supset \{x \mid g(x) \leq 0\}$.

Outer linearization is motivated by convexity. For any point x^i, if f is convex,

$$f(x) \geq f(x^i) + \nabla f(x^i)^T (x - x^i); \tag{3.11}$$

so that $\alpha_i = \nabla f(x^i)^T$ and $\beta_i = f(x^i) - \nabla f(x^i)^T x^i$ can yield the form of f^q for outer linearization applied at successive iterates, $x^1,\ldots,x^q$. Similarly, this approach can be used to approximate each constraint $g_i(x) \leq 0$ or other relaxations as in the lift-and-project method.

The value of outer linearization in convex optimization extends from the ability to capture global properties of the objective and constraints using only local information. Complex nonlinear structures can often be rendered with a parsimonious use of linearizations at a relatively small number of points. New methods built on this approach continue to be developed to take advantage of this property. The following two approaches from nonlinear, discrete optimization and from dynamic optimization, respectively, provide examples.

3.3.1 Nonlinear Mixed-Integer Programming Methods

Outer linearization is the basis for the nonlinear approaches described by Duran and Grossmann [16], Fletcher and Leyffer [18], and the extension in Quesada and Grossmann [24] that is implemented in the solver, FilMINT, by Abhishek et al. [1]. The method applies to the general formulation in (3.1) where Y includes integer restrictions (i.e., $Y \subset \mathbb{Z}^p$). The method solves a relaxation on a branch of a branch-and-bound tree defined by bounds $l \leq y \leq u$, as $(\text{NLPR}(l,u))$ given by

$$\min_{x \in X} f(x,y) \quad \text{s. t.} \quad g(x,y) \leq 0,\ l \leq y \leq u. \tag{3.12}$$

This solution provides a lower bound on the given branch $\big($or overall on (3.1) if $Y \subset [l,u]\big)$.

The method then applies outer linearization again using a form of projection by considering sub-problems, $(\text{NLP}(k))$, given by

$$\min_{x \in X} f(x,y^k) \quad \text{s. t.} \quad g(x,y^k) \leq 0. \tag{3.13}$$

Solving $(\text{NLP}(k))$ for x^k (or solving a corresponding feasibility problem) yields the linearization from the gradient information at x^k as in (3.11) for both the objective and constraints. The outer linearization problem, $(\text{MP}(k))$, at iteration k is then defined by

$$\min_{x\in X,\, y\in Y} f^k(x,y) \quad \text{s. t.} \quad g^k(x,y) \le 0, \tag{3.14}$$

and its continuous relaxation on a branch with restriction $[l,u]$ is (CMP(k)) given by

$$\min_{x\in X,\, y\in [l,u]} f^k(x,y) \quad \text{s. t.} \quad g^k(x,y) \le 0. \tag{3.15}$$

The algorithm proceeds by solving (CMP(k)) at a given node of the branch-and-bound tree to obtain (x^k, y^k). If $y^k \in Y$, then (NLP(k)) is solved, the current upper bound is updated if (NLP(k)) is feasible, and additional cuts are added to CMP(k), which is then solved again. If $y^k \notin Y$, then either additional cuts are generated or a new branch is formed.

3.3.2 Outer Approximation for Convex, Dynamic Optimization

The general form of outer linearization for convex optimization extends at least back to Kelley [21]. The methodology also has appeared frequently in the context of dynamic optimization, particularly for stochastic programs with two and more stages (in, e.g., Dantzig and Madansky [13], Van Slyke and Wets [29], and Birge [7]). This principle can also apply to infinite-horizon dynamic programs as described in Birge and Zhao [9] and summarized here.

The goal is not just to find a single optimum but to find an entire value function V^* of the infinite-horizon problem

$$V^*(x) = \min_{y_1, y_2, \cdots} \sum_{t=0}^{\infty} \delta^t c_t(x_t, y_t) \tag{3.16}$$

$$\text{s.t.} \quad x_{t+1} = A_t x_t + B_t y_t + b_t, \quad \text{for } t = 0,1,2,\ldots, \tag{3.17}$$

$$x_0 = x, \tag{3.18}$$

where $0 < \delta^t < 1$ is a discount factor, and the equation, $x_{t+1} = A_t x_t + B_t y_t + b_t$, characterizes the dynamics of the state transition from stage t to $t+1$. The problem may also include random parameters in c_t and the dynamics (in which case, the objective is an expectation functional).

The above problem can be represented as

$$\begin{aligned} &\min_{y_0} \{c_0(x_0,y_0) + \delta \min_{y_1}\{c_1(x_1,y_1) + \delta \min_{y_2}\{c_2(x_2,y_2) + \ldots\}\}\} \\ &\text{s.t.} \quad x_{t+1} = A_t x_t + B_t y_t + b_t, \quad \text{for } t = 0,1,2,\ldots, \\ &\qquad x_0 = x. \end{aligned}$$

The value function V^* defined by (3.16) is a solution of $V = M(V)$, where the mapping M is defined by

$$M(V)(x) = \min_y \{c(x,y) + \delta V(Ax + By + b)\}. \tag{3.19}$$

For the algorithm to find V^*, suppose that the domain (feasible set) is $D^* = \text{dom}(V^*)$, which is compact and polyhedral. The outer linearization for this method progressively refines an approximation V^k of V^*. Unlike the standard outer linearization, however, each new approximation is only based on an approximation $M(V^k)$ and is not necessarily a support of V^*. The algorithm also must sample throughout D^* to converge to V^*. We let

$$V^k(x) = \max\{Q^i x + q^i : i = 1, \ldots, u_k\},$$

for a set of cuts defined by Q^i and q^i as in the definition of f^k and g^k above. We maintain that $V^k \leq V^*$ and continue to iterate as long as there exists $x \in D^*$ such that $M(V^k)(x) > V^k(x)$.

Outer Approximation for Infinite-Horizon Dynamic Programs

1. Initialization: Find a piecewise linear convex function V^0 satisfying $V^0 \leq V^*$. Set $k \leftarrow 0$.
2. If $V^k \geq M(V^k)$, stop, V^k is the solution. Otherwise, find a point $x^k \in D^*$ with $V^k(x^k) < M(V^k)(x^k)$.
3. Find a supporting hyperplane of $M(V^k)$ at x^k, e.g., $Q^{k+1}x + q^{k+1}$.
   Define $V^{k+1}(x) = \max\{V^k(x), Q^{k+1}x + q^{k+1}\}$.
   $k \leftarrow k+1$. Go to Step 2.

3.4 Inner Linearization

Outer linearization relies on forming an outer approximation of a convex function or convex constraint, while inner linearization, as the name suggests, builds the approximation from within the epigraph of the function or within the feasible region defined by the constraints. In this way, inner linearization implies a restricted version of the original problem while outer linearization implies a relaxation.

The basic approach in inner linearization to a problem of the form (3.10) is to search for a solution in the convex hull of a set of candidate points, $x^1, \ldots, x^k$, with variables corresponding to weights $\lambda_1, \ldots, \lambda_k$ on those points. Problem (3.10) then becomes

$$\min_{\lambda \geq 0, \mathbf{1}^T\lambda = 1} \sum_{i=1}^{k} \lambda_i f(x^i). \tag{3.20}$$

Assuming that $g(x^i) \leq 0$ for each $i = 1, \ldots, k$ ensures that $g(x) \leq 0$ for any $x = \sum_{i=1}^{k} \lambda_i f(x^i)$. To define the inner linearization for a given point x, let

$$F^k(x) = \min_{\lambda \geq 0,\, \mathbf{1}^T\lambda = 1,\, x = \sum_{i=1}^{k} \lambda_i x^i} \sum_{i=1}^{k} \lambda_i f(x^i). \tag{3.21}$$

In general, inner linearization methods, such as Dantzig–Wolfe decomposition (Dantzig and Wolfe [14]), solve the restricted problem in (3.20) and then search for a new solution x^{k+1} that optimizes an auxiliary objective to improve the current approximation. This method is also known as *column generation*, which allows the solution of problems with large numbers of variables without explicitly representing all of them, a particularly valuable strategy in integer programming (see, e.g., Barnhart et al. [3] and Wilhelm [30]). The approach also forms the basis of *generalized linear programming* (Dantzig ([12], chapter 24)) to solve convex programs. In addition, inner linearization forms the foundation for more recent methods as well. As examples, I will describe a recent convex optimization method by Bertsekas and Yu [6] and the linear programming approach to approximate dynamic programming, as described by Van Roy and de Farias [15].

3.4.1 Inner and Outer Approximations for Convex Optimization

The *generalized polyhedral approximation algorithm* in Bertsekas and Yu [6] assumes that (3.10) can be written as

$$\min_{(x^0,\ldots,x^m)\in S} \sum_{i=0}^{m} f_i(x^i), \tag{3.22}$$

where S is a sub-space and $x^i \in \Re^{n_i}, i = 1,\ldots,m$. A problem of the form (3.10) might be represented in this way by re-writing all of the constraints $g_i(x) \le 0$ as $f_i(x)$ using the corresponding indicator function that vanishes ($f_i(x) = 0$) when $g_i(x) \le 0$ and has infinite value ($f_i(x) = +\infty$) when $g_i(x) > 0$. The constraints then follow from $S = \{x^i \mid x^i - x^j = 0, \forall i, j\}$.

This method uses a duality result that the convex conjugate (see Rockafellar [25]) $(f^k)^*$ of an outer linearization f^k of a function f is an inner linearization of the convex conjugate f^* of f. With this observation, each iteration of the algorithm identifies a set of primal solutions x^k and a corresponding set of dual solutions λ^k that correspond to sub-gradients of the relevant approximation of each f_i or, equivalently, are points in the approximation of f_i^*. The algorithm partitions $I_0 = \{0,\ldots,m\}$ as $I_0 = I \cup I_{\text{inner}} \cup I_{\text{outer}}$, where I_{inner} will correspond to f_i that are inner linearized and I_{outer} corresponds to f_i that are outer linearized. The iteration is then to solve

$$\min_{(x^0,\ldots,x^m)\in S} \sum_{i\in I} f_i(x^i) + \sum_{i\in I_{\text{inner}}} f_i^k(x^i) + \sum_{i\in I_{\text{outer}}} F_i^k(x^i), \tag{3.23}$$

where the approximations in f^k and F^k are updated on each iteration. The method either obtains a strictly improving inner or outer linearization in I_{inner} and I_{outer}, respectively, or obtains an optimal solution to (3.22).

3.4.2 Linearization in Approximate Dynamic Programming

The goal of this method is to solve for a value function V^* as in (3.16) with a general set of dynamic equations. In this approach, inner approximation is used instead of outer approximation. An approximation in this case can be written as

$$V^k(x) = \min\left\{ \sum_{i=0}^{n_k} \lambda_i V_i^k \;\middle|\; \sum_{i=0}^{n_k} \lambda_i x^i = x, \sum_{i=1}^{n_k} \lambda_i = 1, \lambda \geq 0 \right\}, \tag{3.24}$$

where $V_i^k = V^k(x^i)$ represents an approximation value at x^i that may be generalized. To achieve efficiencies in this approach, the points can be chosen so that x^0 is a centering point and $x = x^0 + \sum_{i=1}^{n_k} \lambda^i (x^i - x^0)$, where each λ^i can be found quickly as, for example, when $x^i - x^0$ corresponds to positive and negative coordinate directions or when $\text{co}\{x^i, i = 1, \ldots, n_k\}$ is a simplex containing x^0 and x and the λ^i values correspond to the unique barycentric coordinates around x^0. If each V^k is conical in the sense that the epigraph of V^k is a cone centered at x^0, then V_i^k can be extended to obtain approximations throughout the space spanned by $\{x^i - x^0\}$ by centering at x^0 and noting that $V^k(x^i) - V^k(x^0)$ is such that $V^k(x^0 + \rho(x^i - x^0)) = V^k(x^0) + \rho(V^k(x^i) - V^k(x^0))$ for any $\rho \geq 0$. Conditions for this conical property arise in linear optimal control problems (see Birge and Takriti [8]), allowing efficient solutions for that structure.

If each $x^i = x^0 + \mathbf{1}_i$ (i.e., a unit increase in the ith coordinate), the values $\phi_i(x) = (x_i - x_i^0)(V^k(x^i) - V^k(x^0)) + V^k(x^0)$ can be used as a general approximation of $V^*(x)$ in the ith coordination direction. The principle in approximate dynamic programming, as, for example, given in Schweitzer and Seidmann [27], is that this form of linearization can be applied by defining each ϕ_i as a *basis function* with a linear or affine form as here or with more general characteristics. The solution procedure then searches for consistent λ_i values to solve the equation $V^k = M(V^k)$, where $V^k(x, \lambda)$ is now given as

$$V^k(x, \lambda) = \sum_{i=0}^{n_k} \lambda_i \phi_i(x). \tag{3.25}$$

As discussed by de Farias and Van Roy [15], this problem can be solved by the (possibly infinite) linear program

$$\min \int_{x \in X} V^k(x, \lambda) \mu(dx) \quad \text{s. t.} \quad M(V^k) \geq V^k, \tag{3.26}$$

where μ is a weighting measure that assigns positive weight on all possible states $x \in \text{dom}\, V^*$. When X is finite, de Farias and Van Roy show that the error in using this approximation can be bounded by a multiple (that depends on the discount factor and weighting measure) of the error in the best approximation V^k to V^* for any λ. The approximations can also be improved to obtain convergence by suitably choosing the set of ϕ_i functions (see Adelman and Klabjan [22]).

3.5 Conclusions

Geoffrion's [19] description of fundamental problem manipulations for solving large-scale mathematical programs gave a framework for ongoing research to develop increasingly efficient methods to apply to ever wider domains of application. Geoffrion's clear description and unifying treatment of those ideas has provided many subsequent researchers with the insight to make improvements to previous approaches and to uncover new possibilities. The themes of projection, outer linearization, and inner linearization in that paper are indeed the basis for many of the optimization methods that have been proposed since Geoffrion [19] appeared. In this chapter, I have attempted to describe a few of the more recent developments that have built on those fundamental themes. Those fundamental ideas and Geoffrion's clear articulation of them will without a doubt continue to provide researchers with inspiration and guidance for many years to come.

References

1. Abhishek K, Leyffer S, Linderoth JT (2008) FilMINT: An outer-approximation-based solver for nonlinear mixed integer programs. Argonne National Laboratory, Mathematics and Computer Science Division Preprint ANL/MCS-P1374-0906, March 28
2. Balas E, Ceria S, Cornuéjols G (1993) A lift-and-project cutting plane algorithm for mixed 0–1 programs. Mathematical Programming 58:295–324
3. Barnhart C, Johnson EL, Nemhauser GL, Savelsbergh MWP, Vance PH (1998) Branch and price: Column generation for solving huge integer programs. Operations Research 46:316–329
4. Bellman R (1957) Dynamic programming. Princeton University Press, Princeton, NJ
5. Benders JF (1962) Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik 4:238–252
6. Bertsekas DP, Yu H (2009) A unifying polyhedral approximation framework for convex optimization. MIT Working Paper: Report LIDS–2820, September
7. Birge JR (1985) Decomposition and partitioning methods for multi-stage stochastic linear programs. Operations Research 33:989–1007
8. Birge JR, Takriti S (1998) Successive approximations of linear control models. SIAM Journal of Control Optimization 37:165–176
9. Birge JR, Zhao G (2007) Successive linear approximation solution of infinite-horizon dynamic stochastic programs. SIAM Journal of Optimization 18:1165–1186
10. Bixby RE (2002) Solving real-world linear programs: A decade and more of progress. Operations Research 50:3–15
11. Dantzig GB (1959) On the status of multistage linear programming problems. Management Science 6:53–72
12. Dantzig GB (1963) Linear programming and extensions. Princeton University Press, Princeton, NJ
13. Dantzig GB, Madansky A (1961) On the solution of two–stage linear programs under uncertainty. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA
14. Dantzig GB, Wolfe P (1960) The decomposition principle for linear programs. Operations Research 8:101–111
15. de Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. Operations Research 51:850–865

16. Duran MA, Grossmann I (1986) An outer-approximation algorithm for a class of mixed integer nonlinear programs. Mathematical Programming 36:307–339
17. Fiacco Av, McCormick GP (1964) The sequential unconstrained minimization technique for nonlinear programing, a primal-dual method. Management Science 10:360–366
18. Fletcher R, Leyffer S (1994) Solving mixed integer nonlinear programs by outer approximation. Mathematical Programming 66:327–349
19. Geoffrion AM (1970) Elements of large-scale mathematical programming. Management Science 16:652–675
20. Karmarkar N (1984) A new polynomial-time algorithm for linear programming. Combinatorica 4:373–395
21. Kelley JE (1960) The cutting plane method for solving convex programs. Journal of SIAM 8:703–712
22. Klabjan D, Adelman D (2007) An infinite dimensional linear programming algorithm for deterministic semi-Markov decision processes on Borel spaces. Mathematics of Operations Research 32:528–550
23. Lovász L, Schrijver A (1991) Cones of matrices and set-functions and 0-1 optimization. SIAM Journal of Optimization 1:166–190
24. Quesada I, Grossmann IE (1992) An LP/NLP based branch-and-bound algorithm for convex MINLP optimization problems. Computers and Chemical Engineering 16:937–947
25. Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton, NJ
26. Rosen JB (1963) Convex partition programming. In: Graves RL, Wolfe P (eds) Recent advances in mathematical programming. McGraw-Hill, New York, NY
27. Schweitzer P, Seidmann A (1985) Generalized polynomial approximations in Markovian decision processes. Journal of Mathematical Analysis Applications 110:568–582
28. Sherali H, Adams W (1990) A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. SIAM Journal of Discrete Mathematics 3:411–430
29. Van Slyke R, Wets RJ-B (1969) L-shaped linear programs with application to optimal control and stochastic programming. SIAM Journal of Applied Mathematics 17:638–663
30. Wilhelm WE (2001) A technical review of column generation in integer programming. Optimization and Engineering 2:1573–2924

Chapter 4
Multicommodity Distribution System Design by Benders Decomposition* † ‡

A. M. Geoffrion, G. W. Graves§

Abstract A commonly occurring problem in distribution system design is the optimal location of intermediate distribution facilities between plants and customers. A multicommodity capacitated single-period version of this problem is formulated as a mixed integer linear program. A solution technique based on Benders Decomposition is developed, implemented, and successfully applied to a real problem for a major food firm with 17 commodity classes, 14 plants, 45 possible distribution center sites, and 121 customer zones. An essentially optimal solution was found and proven with a surprisingly small number of Benders cuts. Some discussion is given concerning why this problem class appears to be so amenable to solution by Benders' method, and also concerning what we feel to be the proper professional use of the present computational technique.

A. M. Geoffrion, G. W. Graves
University of California, Los Angeles, CA, USA

† Received August 15, 1973.

‡ An earlier version of this paper was presented at the NATO Conference on Applications of Optimization Methods for Large-Scale Resource Allocation Problems, Elsinore, Denmark, July 5–9, 1971. This research was partially supported by the National Science Foundation under Grant GP-36090X and the Office of Naval Research under Contract N00014-69-A-0200-4042. Reproduction in whole or in part is permitted for any purpose of the United States Government.

§ We wish to express our gratitude to Mr. Steven M. Niino, Manager of Operations Research at Hunt-Wesson Foods, Inc., for his outstanding contribution to the success of the practical application reported in this paper. We also want to thank Mr. Shao-Ju Lee of California State University at Northridge for his invaluable assistance in carrying out a difficult computer implementation.

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_4,

4.1 Introduction

4.1.1 The Model

The simplest version of the problem to be modeled is this. There are several commodities produced at several plants with known production capacities. There is a known demand for each commodity at each of a number of customer zones. This demand is satisfied by shipping via regional distribution centers (abbreviated DC), with each customer zone being assigned exclusively to a single DC. There are lower as well as upper bounds on the allowable total annual throughput of each DC. The possible locations for the DC's are given, but the particular sites to be used are to be selected so as to result in the least total distribution cost. The DC costs are expressed as fixed charges (imposed for the sites actually used) plus a linear variable charge. Transportation costs are taken to be linear.

Thus the problem is to determine which DC sites to use, what size DC to have at each selected site, what customer zones should be served by each DC, and what the pattern of transportation flows should be for all commodities. This is to be done so as to meet the given demands at minimum total distribution cost subject to the plant capacity and DC throughput constraints. There may also be additional constraints on the logical configuration of the distribution system.

The mathematical formulation of the problem uses the following notation.

i index for commodities,
j index for plants,
k index for possible distribution center (DC) sites,
l index for customer demand zones,
S_{ij} supply (production capacity) for commodity i at plant j,
D_{il} demand for commodity i in customer zone l,
$\underline{V}_k, \bar{V}_k$ minimum, maximum allowed total annual throughput for a DC at site k,
f_k fixed portion of the annual possession and operating costs for a DC at site k,
v_k variable unit cost of throughput for a DC at site k,
c_{ijkl} average unit cost of producing and shipping commodity i from plant j through DC k to customer zone l,
x_{ijkl} a variable denoting the amount of commodity i shipped from plant j through DC k to customer zone l,
y_{kl} a 0–1 variable that will be 1 if DC k serves customer zone l, and 0 otherwise
z_k a 0–1 variable that will be 1 if a DC is acquired at site k, and 0 otherwise.

The problem can be written as the following mixed integer linear program.

$$\underset{x\geq 0;\, y,\, z=0,\,1}{\text{Minimize}} \sum_{ijkl} c_{ijkl} x_{ijkl} + \sum_{k} \left[f_k z_k + v_k \sum_{il} D_{il} y_{kl} \right] \tag{1}$$

subject to

$$\sum_{kl} x_{ijkl} \leq S_{ij}, \quad \text{all } ij \tag{2}$$

$$\sum_{j} x_{ijkl} = D_{il} y_{kl}, \quad \text{all } ikl \tag{3}$$

$$\sum_{k} y_{kl} = 1, \quad \text{all } l \tag{4}$$

$$\underline{V}_k z_k \leq \sum_{il} D_{ilykl} \leq \bar{V}_k z_k, \quad \text{all } k \tag{5}$$

$$\text{Linear configuration constraints on } y \text{ and/or } z. \tag{6}$$

The notation $y, z = 0, 1$ means that every component y_{kl} and z_k must be zero or one. It is understood that all summations run over the allowable combinations of the indices, since many combinations are either physically impossible (such as an ij combination which signifies a commodity that cannot be made at plant j) or so obviously uneconomical as not to merit inclusion in the model (such as a kl combination that would serve customers in Miami from a DC in Seattle).

The correspondence between this model and the verbal problem statement should be apparent. The quantity $\sum_{il} D_{il} y_{kl}$ is interpreted as the total annual throughput of the kth DC. Constraints (2) are the supply constraints, and (3) stipulates both that legitimate demand must be met (when $y_{kl} = 1$) and that x_{ijkl} must be 0 for all ij when $y_{kl} = 0$. Constraints (4) specify that each customer zone must be served by a single DC. Besides keeping the total annual throughput between $\underline{V}_k$ and $\bar{V}_k$ or at 0 according to whether or not a DC is open, (5) also enforces the correct logical relationship between y and z (i.e., $z_k = 1 \iff y_{kl} = 1$ for some l). Constraints (6) are deliberately not spelled out in detail for the sake of notational simplicity. The only requirement is that they be linear and do not involve any x-variables.

4.1.2 Discussion of the Model

There are several features of the model which warrant some discussion either to point out the flexibility they afford or to indicate the manner in which they differ from related models to be found in the literature.

The reader may have noticed that the transportation variables are quadruply subscripted, whereas previous intermediate location models (Bartakke et al. [2]; Ellwein and Gray [8, p. 296]; Elson [9]; Marks, Liebman and Bellmore [19]) employ separate transportation variables for plant-to-DC and DC-to-customer shipments. That is, we might have used two sets of triply subscripted variables (x_{ijk} and x_{ikl}, say) linked by a flow conservation constraint for each commodity-DC combination. This alternative suffers from a lack of flexibility for some applications because it "forgets" the origin of a commodity once it arrives at a DC. In the real application which sired the work reported in this paper, for instance, the so-called "storage-in-transit" privilege was a very important determinant of rail transportation costs for several of the commodities. A transit rate is figured as the direct plant-customer rate

plus a nominal charge for stopping over at the DC which serves the customer, so long as this DC is not too far off the direct line. The transit rate is usually smaller than the simple sum of the plant-DC rate and the DC-customer rate. Obviously the x_{ijk} and x_{ikl} formulation cannot cope with the transit feature. Another advantage of the x_{ijkl} formulation over the x_{ijk} & x_{ikl} formulation arises when some commodities are perishable; it may be necessary to disallow the possibility of shipping such commodities over jkl routes for which the total journey times are likely to be excessive.

The quadruply subscripted transportation variables also make it easy to accommodate direct plant-customer zone shipments so long as a customer zone does not try to receive a given commodity both from a DC and a plant. For instance, suppose that a certain subset of customer zones is to obtain all commodities directly from the plants instead of via DC's. Then one simply adds a fictitious DC site k_0, say, with the associated z_{k_0} and $y_{k_0 l}$'s fixed at unity, and specifies the rates $c_{ijk_0 l}$ appropriately for each associated ijl (there is no need for (5) to include a constraint for k_0). One may also accommodate the situation in which a customer zone obtains some commodities directly from the plants and the others through its DC. Just make the c_{ijkl}'s corresponding to the direct commodities independent of the possible DC's for such a customer zone, and omit the il combinations corresponding to the directly shipped commodities from both $\sum_{il} D_{il} y_{kl}$ terms in the model.

Another unique feature of the model is that no customer zone is allowed to deal with more than one DC, since the y_{kl}'s must be 0 or 1 and not fractional. Thus each customer's demands must be satisfied by a single DC or directly from a producing plant (as described above). This assumption, which is required by the decomposition technique developed below, is frequently justified in practice. Our first-hand experience with three firms, each in a different industry, is that their accounting systems and marketing structures are geared to serving each customer zone from a single DC. Any change in this convention would be expensive both in terms of added administrative costs and in terms of less convenient service as perceived by customers. There would also be economic disadvantages due to reduced economies of scale in DC-to-customer shipments. Evidently a similar situation exists for other firms, as the desirability of this feature is frequently mentioned by other authors with practical experience [2], [6], [9], [10].

Notice that lower bounds as well as the customary upper bounds may be stipulated on warehouse throughput. This is useful for its own sake when there are reasons why each DC must be larger than a certain minimum size, and also to facilitate using a simple trick to permit a piecewise-linear representation of economies of scale and other nonlinearities (or even discontinuities) in DC costs as a function of throughput: simply introduce alternative DC's at a given site with different size ranges controlled by $\underline{V}_k$ and $\bar{V}_k$, with f_k and v_k specialized accordingly. For instance, a piecewise-linear DC cost function with three pieces would require three alternative DC's (small, medium and large) each with f_k and v_k dictated by the corresponding piece of the DC cost function. A simple configuration constraint can be included among (6) to ensure that at most one of the alternative DC's is opened at each site if this is not an automatic economic consequence of the model. The same trick also

allows some economies-of-scale in transportation costs to be incorporated. This is especially useful for the in-bound (plant-to-DC) component of transportation costs for nontransit commodities. The larger the size range of an alternative DC, the lower should be the unit in-bound rates. The annual throughput of a DC has a much smaller influence on economies-of-scale for the out-bound rates, because the mode of transportation and delivery requirements are largely determined by the customers. This is especially true in view of the model assumption that each customer zone must be supplied by a single DC (the degree of consolidation of out-bound shipments is therefore relatively predictable for a given DC-customer zone pair).

The arbitrary configuration constraints (6) give the model quite a lot of flexibility to incorporate many of the complexities and idiosyncrasies found in most real applications. For instance, (6) permits:

- upper and/or lower bounds on the total number of open DC's allowed;
- specification of subsets of DC's among which at most one, at least one, exactly two, etc., are required to be open;
- precedence relations pertaining to the open DC's (not A unless B, etc.);
- mandatory service area constraints (if DC A is open, it must serve customer zone B);
- more detailed capacity constraints on the size of a DC than (5) permits, as by weighting the capacity consumption characteristics of each commodity differently or by writing separate constraints for individual or subsets of commodities;
- constraints on the joint capacity of several DC's if they share common resources or facilities;
- customer service constraints like

$$\left(\sum_{kl} t_{ikl} D_{il} y_{kl}\right) \Big/ \sum_{l} D_{il} \leq T_i,$$

where t_{ikl} is the average time to make a delivery of commodity i to customer zone l after receiving an order at DC k, and T_i is a desired bound on the average delivery delay for commodity i.

A few additional remarks are in order concerning how the present model fits into the existing literature. Its chief ancestors are, of course, the well-known and much simpler "plant location" models (see Balinski and Spielberg [1, p. 268ff.]; Gray [16]; Ellwein [7] for surveys). These are basically single commodity transportation problems with fixed charges for the use of a source. Often the sources are assumed to have unlimited capacity. Recent work on capacitated problems of this type includes Davis and Ray [5], Ellwein and Gray [8], Fieldhouse [10], Geoffrion and McBride [13], Khumawala and Akinc [17], and Soland [21]. These authors all use branch-and-bound, which has emerged clearly as the most practical optimizing approach.

A natural extension of the capacitated plant location problem to the optimal location of intermediate facilities in multi-echelon systems has been studied by Marks, Liebman and Bellmore [19]. They report reasonably good computational experience with a conventional branch-and-bound algorithm in which the linear programs,

which specialize to capacitated trans-shipment problems, are solved by an out-of-kilter routine. The same model is considered very briefly by Ellwein and Gray [8], who indicate that their capacitated plant location algorithm can be generalized to this case but give no computational experience.

If we now add the multicommodity feature, there appears to be no existing literature on special purpose optimizing algorithms. The only studies of multicommodity intermediate facilities location problems of which we are aware have used general purpose mixed integer linear programming systems. Bartakke et al. [2] describe an application of Bonner and Moore's Functional Mathematical Programming System for the Univac 1108 to an industrial problem with 4 plants, 4 commodities, 10 intermediate distribution sites with 3 possible sizes for each, and 39 customer points. It reportedly required 45 minutes of CPU time to optimize the resulting model with 210 rows, 30 binary variables and 1600 continuous variables. Elson [9] describes a specialized matrix generator and report writer for use in conjunction with the OPHELIE MIXED system for multicommodity intermediate location problems. Computational experience is given for one relatively small problem. The author refers to other computational experience with problems of similar size, from which he estimates that problems with 15 plants, 3 commodities, 45 DC sites, and 50 customer zones can be solved in about $8\frac{1}{2}$ system minutes on the CDC 6600 (assuming a 3:1 conversion ratio of billable system time to central processor time).

The reader who wishes to delve into the literature more deeply is encouraged to consult the excellent and massive (273 page) annotated bibliography on location-allocation systems prepared recently by Lea [18].

4.1.3 Plan of the Paper

§2 specializes Benders' well-known partitioning procedure to our problem in such a way that the multicommodity LP subproblem decomposes into as many independent classical transportation problems as there are commodities. This decomposition makes it possible to solve problems with virtually any number of commodities. Possible points of interest in this section include the technique used to recover the optimal multipliers for each LP subproblem from its analytically reduced and separated components, a variation of Benders' original procedure which has proven effective in this context, and some remarks on the reoptimization capability of this approach via the use of previously generated Benders constraints for revised problems.

§3 briefly describes a full-scale computational implementation which we have used to redesign the national distribution system of a major food firm. This application is discussed at some length in §4, with considerable stress placed on the importance of certain types of pre- and postoptimality runs to the professional success of this study. Actual computational experience is quoted in detail. The reader will be surprised, as we were, that in every run just a few iterations of Benders'

procedure sufficed to find and verify a solution optimal to within a few tenths of one percent. Since this was also true for another large (unrelated) practical problem, it would seem that the class of problems studied herein is unusually amenable to solution by Benders' method.

§5 passes along a lesson learned from early computational experience concerning alternative logically equivalent model representations which are really not equivalent at all when solved by Benders Decomposition. We found that the representation used here is far superior to the natural more compact one we had tried earlier. This phenomenon is examined and implications emerge which may well be useful in other applications of Benders' method.

Some conclusions from our experience to date are offered in §6.

4.2 Application of Benders Decomposition

Most real-life applications of problem (1)–(6) are too large to be solved economically by existing general mixed integer linear programming codes [12]. The application addressed below had 11,854 rows, 727 binary variables and 23,513 continuous variables. The model does, however, have a conspicuous special property that enables it to be decomposed in such a way that the multicommodity aspect becomes much less burdensome: when the binary variables are temporarily held fixed so as to satisfy (4)–(6), the remaining optimization in x separates into as many independent classical transportation problems as there are commodities. This can be seen either from the physical interpretation of the problem or directly from (1)–(3). The transportation problem for the ith commodity is of the form

$$\begin{aligned} &\text{Minimize} \sum_{jl} c_{ij\bar{k}(l)l} x_{ij\bar{k}(l)l} \\ \text{subject to} \quad &\sum_{l} x_{ij\bar{k}(l)l} \le S_{ij}, \quad \text{all } j \\ &\sum_{j} x_{ij\bar{k}(l)l} = D_{il}, \quad \text{all } l \\ &x_{ij\bar{k}(l)l} \ge 0, \quad \text{all } jl, \end{aligned} \tag{7i}$$

where $\bar{k}(l)$ is defined, for each l, as the k-index for which $y_{kl} = 1$ in the temporarily fixed y-array (by (4), $\bar{k}(l)$ is unique for each l).

The simplicity of the problem for fixed (y,z) suggests the application of Benders Decomposition [4]. A conventional specialization of this approach is given in §2.1, and the following section explains how the necessary multipliers of the full subproblem may be analytically synthesized from the multipliers of the reduced and separated subproblems (7i). §2.3 describes a variant of Benders' approach which we have found to be more suitable for computational purposes. Finally, the cost-saving reoptimization capability inherent in this approach is pointed out in §2.4.

4.2.1 Specialization of Benders Decomposition

Application of Benders Decomposition to (1)–(6) in the standard fashion leads to the following algorithm.

Step 0. Select a convergence tolerance parameter $\varepsilon \geq 0$. Initialize $UB = \infty$, $LB = -\infty$, $H = 0$. If a binary airay (y^1, z^1) satisfying (4), (5) and (6) is given, go to Step 2; otherwise, go to Step 1.

Step 1. Solve the current master problem

$$\underset{y,z=0,1;y_0}{\text{Minimize}} \sum_k \left[f_k z_k + v_k \sum_{il} D_{il} y_{kl} \right] + y_0 \tag{8}$$

subject to (4), (5), (6) and

$$y_o + \sum_{ikl} \pi^h_{ikl} D_{il} y_{kl} \geq - \sum_{ij} u^h_{ij} S_{ij}, \quad h = 1, \ldots, H \tag{9}$$

by any applicable algorithm. Let $(y^{H+1}, z^{H+1}, y^{H+1})$ be any optimal solution. Put *LB* equal to the optimal value of (8), which is a lower bound on the optimal value of (1)–(6). Terminate if $UB \leq LB + \varepsilon$.

Step 2.

(a) Solve the linear programming subproblem

$$\underset{x \geqq 0}{\text{Minimize}} \sum_{ijkl} c_{ijkl} x_{ijkl} \tag{10}$$

subject to (2) and (3)
with $y = y^{H+1}$ by any applicable algorithm. Denote the optimal value by $T(y^{H+1})$ and the optimal solution by x^{H+1}. Then the quantity

$$\sum_k \left[f_k z^{H+1}_k + v_k \sum_{il} D_{il} y^{H+1}_{kl} \right] + T(y^{H+1}) \tag{11}$$

is an upper bound on the optimal value of (1)–(6). If (11) is less than *UB*, replace *UB* by this quantity, store $(y^{H+1}, z^{H+1}, x^{H+1})$ as the Incumbent, and terminate if $UB \leq LB + \varepsilon$.

(b) Determine an optimal dual solution for (10) with $y = y^{H+1}$: denote it by u^{H+1} (corresponding to (2)) and π^{H+1} (corresponding to (3)). Increase H by 1 and return to Step 1.

A few remarks on this procedure are in order. First, note that an ε-optimal termination criterion has been used. The available upper and lower bounds on the optimal value of (1)–(6) coincide to within ε upon termination, at which time the Incumbent has been demonstrated to be ε-optimal in (1)–(6). Prior to termination it is known only that the Incumbent is within $(UB - LB)$ of the optimal value. Finite convergence is assured for any $\varepsilon \geq 0$.

Second, note that no provision is made at Step 2 for the possibility that (10) may be infeasible for some choices of y. This possibility can be handled easily within the standard framework of Benders Decomposition by slightly complicating the above algorithm, but we elect to preclude it here by assuming without loss of generality that $\sum_j S_{ij} \geq \sum_l D_{il}$ for all i (otherwise (1)–(6) is infeasible) and that all possible jk combinations are technically allowed (if j_0k_0 corresponds to an uneconomical route, take $c_{ij_0k_0l}$ equal to any comparatively large number). It is not difficult to verify that these innocuous assumptions imply that (10) is feasible and has a finite optimal solution for every binary y satisfying (4).

Third, as indicated previously, the LP subproblem (10) is most easily solved by solving an equivalent collection of independent classical transportation problems—one for each commodity. This can be demonstrated by observing that since y^{H+1} satisfies (4), (3) implies

$$x_{ijkl}^{H+1} = 0 \quad \text{for all } ijkl \text{ with } k \neq \bar{k}(l)$$

where $\bar{k}(l)$ is the k-index for which $y_{kl}^{H+1} = 1$. Thus (10) simplifies to

$$\text{Minimize} \sum_i \left(\sum_{jl} c_{ij\bar{k}(l)l} x_{ij\bar{k}(l)l} \right)$$

subject to

$$\sum_l x_{ij\bar{k}(l)l} \leq S_{ij}, \quad \text{all } ij$$

$$\sum_j x_{ij\bar{k}(l)l} = D_{il}, \quad \text{all } il$$

$$x_{ij\bar{k}(l)l} \geq 0, \quad \text{all } ijl.$$

This problem obviously separates on i into independent transportation problems of the form (7i). If the optimal value of (7i) is denoted by $T_i(y^{H+1})$, then $T(y^{H+1}) = \sum_i T_i(y^{H+1})$.

The reduction of (10) to independent problems of the form (7i) greatly simplifies Step 2a, but Step 2b then becomes less straightforward. The required optimal dual solution for (10) must be synthesized from the optimal dual solutions of (7i). The relationship between the optimal primal solutions of (10) and (7i) is obvious, but the relationship between the optimal dual solutions requires some analysis. This analysis is as follows.

4.2.2 Details on Step 2b

Step 2b requires an optimal dual solution (u^{H+1}, π^{H+1}) to (10) with y fixed at y^{H+1}. Since (10) is solved via (7i) rather than directly, the required dual solution must be synthesized from the available dual optimal solutions to (7i).

For notational simplicity, the superscript $H+1$ will be replaced by an overbar (e.g., y^{H+1} becomes $\bar{y}$). Denote the available optimal dual variables of (7i) by $\bar{u}_{ij}$ (corresponding to the supply constraints) and $\bar{v}_{il}$ (corresponding to the demand constraints). It will be shown that the appropriate formulae to be used at Step 2b are:

$$\bar{u}_{ij} = \bar{\mu}_{ij}, \quad \text{all } ij \tag{12a}$$

$$\bar{\pi}_{ikl} = \operatorname*{Max}_{j}\{-\bar{\mu}_{ij} - c_{ijkl}\}, \quad \text{all } ikl. \tag{12b}$$

To derive (12), one must compare the duals of (10) with those of (7i), where y is fixed at $\bar{y}$. The dual of (10) is

$$\begin{aligned} \operatorname*{Maximize}_{u\geqq 0;\,\pi} & \sum_{ikl} \pi_{ikl}(-D_{il}\bar{y}_{kl}) + \sum_{ij} u_{ij}(-S_{ij}) \\ \text{subject to} & \\ & -u_{ij} - \pi_{ikl} \leq c_{ijkl}, \quad \text{all } ijkl. \end{aligned} \tag{13}$$

Notice that for any fixed u, the optimal choice of π is obvious since there are no joint constraints on π and each π_{ikl} is constrained only from below by the bound

$$b_{ikl}(u) \triangleq \operatorname*{Max}_{j}\{-u_{ij} - c_{ijkl}\}.$$

If $(-D_{il}\bar{y}_{kl}) < 0$ then the best choice of π_{ikl} is $b_{ikl}(u)$, while if $(-D_{il}\bar{y}_{kl}) = 0$ then the optimal choice is any number greater than or equal to $b_{ikl}(u)$.

Notice also that when $(-D_{il}\bar{y}_{kl}) = 0$, as when $k \neq \bar{k}(l)$, the corresponding constraints may simply be dropped from (13) since they may always be satisfied without any effect on the value of the objective function. Thus (13) is equivalent to

$$\begin{aligned} \operatorname*{Maximize}_{u\geqq o;\,\pi_{i\bar{k}(l)l},\,\forall il} & \sum_{il} \pi_{i\bar{k}(l)l}(-D_{il}\bar{y}_{\bar{k}(l)l}) + \sum_{ij} u_{ij}(-S_{ij}) \\ \text{subject to} & \\ & -u_{ij} - \pi_{i\bar{k}(l)l} \leq c_{ij\bar{k}(l)l}, \quad \text{all } ijl, \end{aligned} \tag{14}$$

with the understanding that for ikl with $k \neq \bar{k}(l)$, $\bar{\pi}_{ikl}$ is any number greater than or equal to $b_{ikl}(\bar{u})$.

Now consider the duals of (7i) for each i, which may be combined into a single linear program since there are no variables in common. That is, $(\bar{\mu}, \bar{v})$ is an optimal solution of

$$\begin{aligned} \operatorname*{Maximize}_{\mu\geqq 0;\,v} & \sum_{i} \left[\sum_{j} \mu_{ij}(-S_{ij}) + \sum_{l} v_{il}(-D_{il}) \right] \\ \text{subject to} & \\ & -\mu_{ij} - v_{ij} \leq c_{ij\bar{k}(l)l}, \quad \text{all } ijl. \end{aligned} \tag{15}$$

Comparison of (14) and (15) reveals that these are identical optimization problems (remember that $\bar{y}_{\bar{k}(l)l} = 1$), and hence the choice

$$\bar{u}_{ij} = \bar{u}_{ij}, \quad \text{all } ij \tag{16a}$$

$$\pi_{i\bar{k}(l)l} = \bar{v}_{il}, \quad \text{all } il \tag{16b}$$

is optimal in (14). In view of the previous discussion, we also have the following necessary (given (16a)) and sufficient condition on the remaining $\bar{\pi}_{i\bar{k}l}$'s:

$$\bar{\pi}_{ikl} \geq \operatorname*{Max}_j \{-\mu_{ij} - c_{ijkl}\}, \quad \text{for all } ikl \text{ with } k \neq \bar{k}(l). \tag{16c}$$

Relations (16a)–(16c) give the desired complete optimal solution to (13). Since (16a) is identical to (12a), it remains but to reduce (16b) and (16c) to the form (12b).

Relation (16c) is easily converted to the form of (12b) by selecting $\bar{\pi}_{ikl}$ in (16c) to be as small as possible, that is, so that equality holds—for, by the nonnegativity of $D_{il}y_{kl}$ in (9), this gives the best approximation to the optimal transportation cost function T. Second, by inspection of (15) we see that

$$\bar{v}_{il} = \operatorname*{Max}_j \{-\bar{\mu}_{ij} - c_{ij\bar{k}(l)l}\}, \quad \text{all } il\text{: } -D_{il} < 0 \tag{17a}$$

$$\bar{v}_{il} \geq \operatorname*{Max}_j \{-\bar{\mu}_{ij} - c_{ij\bar{k}(l)l}\}, \quad \text{all } il\text{: } -D_{il} = 0. \tag{17b}$$

We may assume without loss of generality that equality holds in (17b), for if not then one may simply redefine $\bar{v}_{il}$ so that it does hold without upsetting the optimality of $(\bar{\mu}, \bar{v})$ in (15). Hence

$$\bar{v}_{il} = \operatorname*{Max}_j \{-\bar{\mu}_{ij} - c_{ij\bar{k}(l)l}\}, \quad \text{all } il,$$

which shows that (16b) reduces to (12b) and concludes the proof of (12).

4.2.3 The Variant Actually Used

There are numerous variants of the pure Benders Decomposition algorithm described in §2.1. One variant of particular interest is not to solve the current master problem at Step 1 to optimality, but rather to stop as soon as a feasible solution to it is produced which has value below $UB - \varepsilon$. This implies, of course, that the master problem no longer produces a lower bound on the optimal value of (1)–(6) and so LB must be inactivated. The termination criterion of Step 2a must be deleted and that of Step 1 must be replaced by: "terminate if the current master problem has no feasible solution with value below $UB - \varepsilon$; the current Incumbent is an ε-optimal solution of (1)–(6)."

It is not difficult to see that this variant must converge to an ε-optimal solution within a finite number of iterations. This follows from the finiteness of the number of dual solutions of (10) and from the easily verified fact that if any dual solution should be produced more than once at Step 2b, then the Incumbent must be

improved by at least ε at each such repetition. There can be no more than a finite number of repetitions because the optimal value of (1)–(6) is bounded below.

The principal motivation behind this variant is that the early master problems have too little information about transportation costs to be worth optimizing very strictly. It takes several "Benders cuts" of the form (9) in order to give accurate information concerning these costs. This suggests that the master problems should be *sub*optimized, particularly when H is small. The degree of optimality achieved by this variant increases with H for two reasons: the minimal value of the master problem increases as H increases due to the accumulation of cuts, and the threshold $UB-\varepsilon$ decreases each time an improved Incumbent is found.

A second motivation is that the variant's master problems are feasibility-seeking only:

$$\text{Find } y, z = 0, 1 \quad \text{and} \quad y_0 \text{ to satisfy } (4), (5), (6), (9) \text{ and}$$
$$\sum_k \left[f_k z_k + v_k \sum_{il} D_{il} y_{kl} \right] + y_0 \le UB - \varepsilon$$

or, equivalently upon elimination of y_0,

$$\text{Find } y, z = 0, 1 \text{ to satisfy } (4), (5), (6) \text{ and}$$
$$\sum_k \left[f_k z_k + v_k \sum_{il} D_{il} y_{kl} \right] - \sum_{ij} u_{ij}^h S_{ij} - \sum_{ikl} \pi_{ikl}^h D_{il} y_{kl} \le UB - \varepsilon, \qquad h = 1, \ldots, H. \tag{9a}$$

Thus we may literally introduce any appealing (linear) objective function, say $\phi(y,z)$, and take the master problem to be:

$$\underset{y,z=0,1}{\text{Minimize}}\, \phi(y,z) \quad \text{subject to (4), (5), (6) and (9a).} \tag{8a}$$

It is not necessary to optimize (8a), of course, but merely to produce a feasible solution if one exists. The choice of ϕ should be so as to encourage the production of useful feasible solutions. We have found the last (Hth) function appearing on the left-hand side of (9a) to be a good choice in practice.

We remark that (8a) is a pure 0–1 integer program, whereas (8) is a mixed integer program due to the appearance of y_0. This gives (8a) the added advantage of being somewhat more convenient to work with.

4.2.4 Re-Optimization

One of the advantages of the Benders Decomposition approach is that it offers the possibility of making sequences of related runs in considerably reduced computing times as compared with doing each run independently. The need for multiple runs is particularly acute in distribution system design studies because of the great economic consequences of the final solution, the difficulties of ascertaining

future demands and costs with precision, and other reasons discussed at some length in §4.2.

The reoptimization capability of Benders' approach is due to the fact that the cuts (9a) generated to solve one problem can often be revised with little or no work so as to be valid in a modified version of the same problem. Assume for a moment that this is so. Then the modified problem can be started with these old (possibly revised) cuts included in the initial master problem and each master thereafter. If the optimal (y,z) solution of the modified problem is not too far from the optimal (y,z) solution of the original problem, then one would expect termination of the procedure in fewer major iterations than would be the case if it were begun from scratch.

The revision of cuts so as to be valid in a modified version of the problem is an easy matter so long as the c_{ijkl} coefficients do not decrease. This limitation is due to the requirement that (u^h, π^h) in (9) and (9a) must be feasible in the dual subproblem (13) corresponding to the modified version. Thus, increasing some c_{ijkl}'s and making arbitrary changes in the $\underline{V}_k$'s and $\bar{V}_k$'s and in the configuration constraints (6) require no revisions at all in (9a); except, of course, that appropriate values of UB and ε must be used. Changing an f_k or v_k is easily accomplished by a simple revision formula. Changing an S_{ij} or D_{il}, on the other hand, requires forethought in that the u_{ij}'s and π_{ikl}'s themselves enter into the revision formulae; normally these duals will not be saved since there is no need for them once a cut is calculated. Saving the u_{ij}'s poses no particular problem because the number of allowable ij combinations is relatively small in most applications. This would permit arbitrary changes in the S_{ij}'s. Saving all of the π_{ikl}'s would be burdensome storage-wise, so it is best to reconstruct them from the u_{ij}'s via (12). Thus arbitrary changes in the D_{il}'s are only slightly more difficult to accommodate than changes in the S_{ij}'s.

The usefulness of this reoptimization capability is indicated by the computational experience presented in §4.3.

4.3 Computer Implementation

An elaborate all-FORTRAN implementation has been carried out for the variant of Benders Decomposition described in §2. The objective of solving large problems in moderate computing times required the use of efficient algorithms for solving the master problems and subproblems, and careful data management techniques. These matters are discussed briefly in this section.

4.3.1 Master Problem

The master problems, of the form (8a), are pure 0–1 integer linear programs with a variable for every allowable DC-customer zone combination (y_{kl}) and for every possible DC site (z_k). Typically this leads to at least several hundred binary

variables. Thus it was necessary to devise a specialized method which exploits the special structure of (8a). The method we employ is a hybrid branch-and-bound/cutting-plane approach with numerous special features.

The cuts employed are the original mixed integer cuts proposed by Gomory in 1960, and are applied to each node problem in order to strengthen the LP bounds and to drive variables toward integer values in preparation for the choice of a branching variable. Absolute priority is given to z-variables over y-variables in branching. Reversal bounds are calculated for variables which are branched upon using relaxed versions of (8a) which drop the integrality requirements on y (while keeping the integrality requirements on z) and transfer a linear combination of all constraints except (5) and individual variable bounds up into the objective function [11]. The multipliers which determine the linear combination are the appropriate dual variables of a node problem solved as a linear program (ignoring the integrality requirements on both y and z).

The linear programming subroutine takes full advantage of the generalized upper bounding constraints (4), and also exploits certain other aspects of the problem structure. It economizes on the use of core storage by generating columns as needed from compactified data arrays.

Finally, it should be mentioned that a number of logical relationships between the variables are built in at various points of the master problem algorithm so as to detect several kinds of infeasibility and "fix" the free variables when this is justified.

4.3.2 Subproblem

The transportation subproblems (7i) are solved using a new primal simplex-based algorithm with factorization (Graves and McBride [15]). Contrary to the conventional wisdom, such methods are superior to out-of-kilter type algorithms for most network flow applications [14], [15]. This is certainly true for the present application, where only the costs of the transportation subproblems change between successive solutions. An earlier implementation using an out-of-kilter algorithm was an order of magnitude slower on the average.

4.3.3 Data Input and Storage

Core storage requirements are economized by extensive use of overlay, cumulative indexing, and the creation of compact data sets from which model coefficients can be generated conveniently as needed. Most of the larger of these data sets are kept on disk. Raw problem data pertaining to permissible $ijkl$ combinations, transportation costs, and customer demands are input from tape to a preprocessor program which creates the appropriate data sets on disk. These are then accessed directly by the main program, which receives the rest of the problem data (S_{ij}, $\underline{V}_k$, $\bar{V}_k$, f_k,

v_k and configuration data for (6)) from direct keyboard input using the URSA conversational CRT-display remote job entry system at UCLA. The editing and scope display facilities of URSA make this an ideal means of entering and revising all but bulk data. Matrix generation and similar chores are accomplished entirely by the preprocessor and main programs.

The specific types of configuration constraints (6) accommodated in the current program include: fixing selected y_{kl} and z_k variables at specific values to set up regional or otherwise reduced versions of the full problem; mutual exclusivity constraints on DC sites; mandatory service area constraints for each DC; and a limit on the maximum number of DC's that may be open.

Newly generated cuts are stored on disk for use in the reoptimization mode described in §2.4. The last primal transportation solution is also stored on disk to serve as an advanced start in subsequent runs for which it is still feasible.

4.4 Solution of a Large Practical Problem

4.4.1 Overview

Hunt-Wesson Foods, Inc., produces several hundred distinguishable commodities at 14 locations (Wesson refineries, Hunt canneries, and co-packers) and distributes nationally through a dozen distribution centers. The firm decided in 1970 to undertake a thorough study of its distribution system design with particular emphasis on the question of distribution center locations. The study was prompted both by the need to resolve several expansion and relocation issues that had arisen, and by the recognition that a systematic global study of the entire distribution system would be likely to disclose opportunities for improvement that could not be identified by conventional analyses of individual cases and geographic regions.

The primary outcome of the study was that five changes were recommended in the firm's configuration of distribution centers (the movement of existing DC's to different cities and the opening of new DC's). The three most urgent of these changes have been carried out as of this writing and the other two are in process. The realizable annual cost savings produced by the study are estimated to be in the low seven figures.

§4.2 describes the various types of computer runs needed to carry out the study. Actual computational experience is summarized in §4.3.

4.4.2 Eight Types of Computer Runs

It is obvious that most distribution system design problems are of sufficiently major economic consequence to warrant the most careful computational treatment. Yet we were surprised by the large number of runs needed to deal properly with the various aspects of a real application. No less than 8 different types of runs can

be distinguished, each of which may require several—sometimes many—distinct submissions:

- probationary exercises
- regional optimization
- global optimization
- "what if . . . ?"
- sensitivity analysis
- continuity analysis
- tradeoff analysis
- priority analysis.

For obvious reasons we cannot go into detail on all of these phases of the study, but we would like to make some general remarks on each in the light of our experience.

The purpose of the probationary exercises is to expose any possible shortcomings of the model, data, or computer code that may compromise their managerial usefulness. They must be regarded as "on probation" until proven otherwise, no matter how meticulous have been the data verification and program debugging efforts. A series of exercises is required in which the computer competes with management in carefully designed decision situations. Each situation must be limited in scope, as by restricting the number of free optimization variables, so that its complexity does not overwhelm the managers' ability to apply experience and familiar analytical techniques—yet it should be broad enough to exercise a significant portion of the model. The computer's solution and the managers' solution must be compared and any significant discrepancies must be reconciled, by hand calculation if necessary. For instance, it is useful to run the problem locked in to the current configuration of distribution centers so that the only optimization required is service area design and transportation flows. The series of exercises should involve each part of the model at least once. In this fashion the model, data and computer code truly earn their credibility.

Regional optimizations focusing on natural geographical regions are bridges between probationary exercises and global optimizations runs, to help tune internal algorithmic parameters and tactics while producing useful results. Four such regions were sufficient in our application.

Far from being the climax of a study, a global optimization run with all decision variables free requires considerable further study to confirm its validity and enhance its usefulness. It spawns additional runs to answer management's many inevitable "what if . . . ?" questions (what if a certain DC were kept open, or a certain customer zone were serviced by another DC, or a better rail rate negotiated here or DC lease there, etc.). It also raises questions concerning the sensitivity of the optimal solution to variation of the data. The need to address such questions is taken for granted in applications of linear programming but they are often slighted in large-scale integer programming applications—presumably on the grounds of their excessive computational cost. Our experience, however, is that such runs are indispensable as a source of useful insight into the behavior of the model and its tolerance for estimation

errors. For instance, they revealed a serious error made during the initial formulation of the model concerning the specification of the lower limits $\underline{V}_k$ on distribution center throughput. Runs done using demand projected for several years beyond the primary target period of the study gave reassurance that dynamic factors were not unduly difficult to cope with via the static model used here.

Continuity analysis is similar to sensitivity analysis except that the purpose is to discover a possible pathology which cannot arise for ordinary linear programming models. We are referring to the possibility that a small change in the data may induce a sudden incommensurately large decrease in the optimal value of (1)–(6), a situation to which a modeler is likely to be quite averse since almost any datum can be changed by a small amount for a commensurately small cost (see Williams [23]). This situation can occur when the data changes lead to a discontinuous change in the feasible region. Changes in data appearing only in the objective function (1) [i.e., in the c_{ijkl}, v_k and f_k coefficients] cannot lead to such behavior. The other data should be checked by doing a run which relaxes each such coefficient somewhat; that is, each $\underline{V}_k$ should be decreased and each $\bar{V}_k$ and S_{ij} should be increased (it can be shown that relaxation of the $\underline{V}_k$'s and $\bar{V}_k$'s precludes the need to perturb the D_{il}'s). If the decrease in the optimal value of (1)–(6) is excessively large by comparison with the estimated economic cost of changing these coefficients,[1] then additional more specific runs must be undertaken to localize the source of difficulty. A managerial decision would then have to be made concerning possible revisions of the problem data or even of the model itself. No serious discontinuities were detected for this application.

Tradeoff analysis runs are appropriate when there are other major quantifiable criteria besides cost in evaluating the desirability of a given distribution system design. Perhaps the most important secondary criterion is the quality of customer service as it depends upon the distance between a DC and the customer zones it serves. One possibility is to adopt the average delivery delay criterion suggested in §1.2 and to solve the problem with successively tighter T_i's. In this manner one may generate the tradeoff curve between total distribution cost and the average delivery delay for any given product or weighted combination of products.

The last type of run on the list is priority analysis. When a study reaches the point where management is ready to consider practical implementation of the results, it is useful to distinguish the aspects of the solution yielding the largest savings from those of relatively marginal significance. Runs done to help refine this distinction suggest which aspects of the solution most urgently call for implementation and which should be postponed or even dropped as too marginal to be worth the organizational upset. In the present application this mainly involved trying to assess the relative economic value of each of the major changes recommended for the distribution center configuration then extant. The actual process is summarized in Table 1, which focuses on the distribution center locations because these are the decisions of primary managerial concern. As the first row indicates, the optimal DC configuration can be viewed as requiring 6 changes to the current (1970) configuration.

[1] Only the coefficient changes actually required for feasibility of the new solution would, of course, enter into this estimation.

Table 1 Priority analysis results

DC Locations		Service Areas and Transport	Total Cost	Differences
OPT Optimum (6 changes)		Optimum	100.00	
A Current		Current	103.15	
B Current		Optimum	101.43	1.72 save over A
B.1	1	"	101.45	−0.02
B.2	2	"	101.34	0.09
B.3 Current & One	3	"	101.14	0.29 save
B.4 Change:	4	"	101.42	0.01 over
B.5	5	"	101.37	0.06 B
B.6	6	"	100.71	0.72
B.7 Current & Best	1	"	100.01	0.01 loss
B.8 Subset of Changes	2	"	100.12	0.12 over
B.9 Omitting:	4	"	100.13	0.13 OPT
C Current & Changes 3,5,6		Optimum	100.30	1.13 save over B
C.1 Current & Changes	1	"	100.29	0.01 save
C.2 3, 5, 6 and Also:	2	"	100.17	0.13 over
C.3 4	4	"	100.13	0.17 C
D Current & Changes 2, 3, 4, 5, 6		Optimum	100.01	0.29 save over C

Some of the changes require relocating an existing DC to a different city and the others require opening a new DC. The total distribution costs corresponding to the optimal configuration are normalized to 100. Row A gives the relative total cost corresponding to the current DC configuration and also the current service areas and transportation flows. Row B retains the current DC configuration but optimizes the service areas and transportation flows. Notice that slightly more than half of the total possible savings could be achieved by service area and transportation flow realignments alone.

Now comes the analysis of the relative value of each of the 6 changes, from which some subset is to be selected for implementation. Runs B.1–B.6 indicate the savings of each change if done individually. Changes 3 and 6 appear to be very attractive, changes 2 and 5 only moderately attractive, and changes 1 and 4 unattractive. Change 5, however, was quite appealing to management on the grounds that it would give additional warehousing space in a region of the country where space was in particularly short supply. Management therefore was inclined to give top implementation priority to changes 3, 5 and 6. This inclination was supported by the results of runs B.7–B.9, which examine the effect of omittinig one of the *other* changes and selecting the best subset of the remainder. It turned out that changes 3, 5 and 6 were among those selected in every case. Top priority was therefore given to changes 3, 5 and 6 which, row C reveals, jointly save a little more than one would expect from simply adding their individual savings (1.13 versus 1.07). Changes 1, 2 and 4 were examined again individually given the acceptance of 3, 5 and 6. Changes 2 and 4 now look quite attractive, while 1 continues to be borderline. This conclusion is supported by runs B.7–B.9 because the same results would have been obtained if changes 3, 5 and 6 had been mandatory in these runs. In light of this analysis and of

factors outside the scope of the model, management gave second priority to changes 2 and 4. Change 1 was considered too marginal for implementation. Row D shows that changes 2 through 6 are only 1/100 of 1% away from the system optimum.

4.4.3 Computational Performance

This section summarizes the code's computational performance on the Hunt-Wesson problem. All computing times refer to UCLA's IBM 360/91.

Table 2 presents ten representative runs without use of the reoptimization technique discussed in §2.4, and three with it (labeled R). None of these runs incorporated any type (6) configuration constraints beyond the locking open or closed of certain distribution center sites, so that the reader would be assured that the problem was not so severely constrained as to greatly facilitate optimization (our experience has been that while configuration constraints do tend to make the problem easier, the influence on computing times is rarely dramatic). Runs 6 and 7 are identical except for the specification of ε. Runs 8 and 9 are identical except that the $\underline{V}_k$'s were all 10% higher in run 8. Runs 2R, 3R and 6R are identical to runs 2, 3 and 6 respectively, except that each was initiated using all of the cuts generated by runs 1, 5 and 4, respectively. The largest number of free DC sites in any of these runs is 30 because the remaining sites were determined to be dominated as obviously uneconomical during the probationary exercises and regional optimizations.

The most striking conclusion to be drawn from Table 2, and indeed from our entire computational experience, is the surprisingly small number of iterations

Table 2 Representative runs

Run No.	DC's		Free 0–1 Variables[a]	Rows	ϵ (%)[b]	Major Iter.	Execution Time (Sec.)[c]
	Free	Locked Open					
1	0	16	249	4,403	0.06	3	16.7
2	0	16	254	4,488	0.03	4	23.8
2R	0	16	254	4,488	0.03	4	16.6
3	7	11	287	4,944	0.03	5	25.5
3R	7	11	287	4,944	0.03	4	17.5
4	15	4	336	5,657	0.06	4	23.2
5	20	1	349	5,783	0.15	4	24.9
6	20	5	411	6,857	0.06	7	50.5
6R	20	5	411	6,857	0.06	5	38.1
7	20	5	411	6,837	0.15	4	29.4
8	25	1	427	7,054	0.15	5	43.8
9	25	1	427	7,054	0.15	5	37.7
10	30	1	513	8,441	0.15	5	191.0

(*Notes*: (a) z_k's corresponding to the free DC's plus y_{kl}'s corresponding to DC's either free or locked open; (b) percentage of the optimal total cost; (c) in addition to execution time, each run required about one second of link editing time.)

required for convergence even with very small values of the optimality tolerance ε. The number of iterations increases only slowly with the size of the problem. Some partial explanations for this fortunate state of affairs are offered in the next section.

Table 3 gives further details on the runs listed in Table 2. For convenience, the optimal value of each run is normalized to 100. The difference between the "total" and "master" columns is the time at each major iteration spent extracting and solving the 17 transportation problems plus cut generation time. About half of this time, which runs quite consistently around 5 seconds, is spent performing the extraction from the data sets on disk.

Table 3 Detailed results for the runs of Table 2

Run No.	Major Iteration	Value of Design from Current Master (11)	Execution Time (Sec.)	
			Master	Total
1	1	103.51	5.8	11.5
	2	100.00	0.2	5.1
	3	Termination	0.1	0.1
2	1	102.78	7.5	13.1
	2	100.00	0.2	5.3
	3	100.01	0.6	5.1
	4	Termination	0.3	0.3
2R	1	100.02	0.9	6.9
	2	100.04	0.2	4.5
	3	100.00	0.3	5.0
	4	Termination	0.2	0.2
3	1	102.96	3.7	9.4
	2	100.04	0.2	5.3
	3	100.01	0.4	5.4
	4	100.00	0.3	5.3
	5	Termination	0.1	0.1
3R	1	100.04	0.9	6.9
	2	100.02	0.4	5.1
	3	100.00	0.5	5.3
	4	Termination	0.2	0.2
4	1	102.00	1.3	6.9
	2	100.01	0.5	5.3
	3	100.00	3.7	8.6
	4	Termination	2.4	2.4
5	1	101.94	1.3	6.8
	2	100.30	0.5	5.5
	3	100.00	5.4	10.4
	4	Termination	2.2	2.2
6	1	102.95	1.4	7.2
	2	100.40	0.6	5.7
	3	100.35	2.5	7.5
	4	100.29	2.6	7.5
	5	100.19	1.1	6.1
	6	100.00	0.3	5.3
	7	Termination	11.2	11.2
6R	1	100.39	0.8	7.3
	2	100.34	0.2	5.0
	3	100.30	5.9	10.8
	4	100.00	0.9	6.0
	5	Termination	9.0	9.0

Table 3 (Continued)

Run No.	Major Iteration	Value of Design from Current Master (11)	Execution Time (Sec.)	
			Master	Total
7	1	102.90	1.5	7.1
	2	100.36	0.5	5.4
	3	100.00	3.3	8.8
	4	Termination	8.1	8.1
8	1	103.86	0.7	7.3
	2	100.37	0.7	5.7
	3	100.15	4.6	9.8
	4	100.00	0.5	5.4
	5	Termination	15.6	15.6
9	1	104.09	1.5	7.0
	2	100.37	0.6	5.6
	3	100.20	2.7	7.8
	4	100.00	0.5	5.5
	5	Termination	11.8	11.8
10	1	105.51	1.6	6.9
	2	100.38	0.5	5.3
	3	100.19	2.7	7.6
	4	100.00	0.3	5.1
	5	Termination	166.1	166.1

From Table 3 it can be seen that suboptimizing the master problem as described in §2.3 is generally successful in helping to keep the time spent on it quite small. As one might expect, the final master problem tends to be relatively difficult for the larger problems. Notice also that the actual cost of the designs produced by the successive master problems usually (but not always) improves monotonely. Finally, we can observe that reoptimization saves computing time but not necessarily major iterations, and that it tends to yield a good first design.

It should be emphasized that the same standard internal and external parameter settings have been used in all of the runs. This was done in the interest of comparability. But, obviously, many useful alternatives exist which may lead to improved performance in specific cases. For instance, gradually reducing ε at each major iteration is a more effective way to achieve a desired low final ε at termination than keeping it constant. Initializing UB at a good known upper bound less than $+\infty$ is also possible and beneficial in most runs. And selectivity in choosing which prior cuts to use for reoptimization is helpful. All such ad hoc adjustments have been avoided here.

4.5 A Lesson on Model Representation

Anyone accustomed to working with linear programming applications is inclined to economize on the number of constraints he uses in a large-scale model. The model (1)–(6) presents an obvious opportunity to economize on the number of type (3)

constraints without changing the logical content of the model in any way: replace (3) by

$$\sum_{jk} x_{ijkl} = D_{il} \quad \text{all } il \tag{3a}$$

$$\sum_{ij} x_{ijkl} = \left(\sum_{i} D_{il}\right) y_{kl}, \quad \text{all } kl. \tag{3b}$$

This formulation performs the two functions of (3) separately, namely ensuring that all demands are met and enforcing the appropriate logical relationship between the x's and the y's. The resulting representation of the problem is equivalent (has the same set of feasible solutions), and usually has fewer constraints. For the Hunt-Wesson application, the representation using (3a) and (3b) in place of (3) has 8,855 fewer constraints!

It turns out, however, that it would be a serious mistake to use this representation with any type of Benders Decomposition approach. The reason is that it leads to much weaker cuts. To see this, recall that all variants of Benders Decomposition work by accumulating linear supports to $T(y)$, which is defined in Sec. 2.1 as the optimal total transportation cost as a function of the configuration design y. For a given binary $\bar{y}$ satisfying (4),

$$-\sum_{ij} \bar{u}_{ij} S_{ij} + \sum_{kl} \left(-\sum_{i} D_{il} \bar{\pi}_{ikl}\right) y_{kl} \tag{18}$$

is such a support, where $\bar{u}$ and $\bar{\pi}$ are defined as in (12). This support is derived from the original formulation of the problem using (3) and is implicit in (9) and (9a). The corresponding support for the revised formulation using (3a) and (3b) in place of (3) can be written as:

$$-\sum_{ij} \bar{u}_{ij} S_{ij} + \sum_{kl} \left[-\sum_{i} D_{il} \left(\bar{\pi}_{i\bar{k}(l)l} + \operatorname*{Max}_{i'} \{\bar{\pi}_{i'kl} - \bar{\pi}_{i'\bar{k}(l)l}\}\right)\right] y_{kl}. \tag{18a}$$

It is evident by inspection (subtract $\bar{\pi}_{i\bar{k}(l)l}$ from both sides) that

$$\bar{\pi}_{ikl} \leq \bar{\pi}_{i\bar{k}(l)l} + \operatorname*{Max}_{i'} \{\bar{\pi}_{i'kl} - \bar{\pi}_{i'\bar{k}(l)l}\} \quad \text{for all } ikl,$$

with the magnitude of the difference increasing with the number of commodity classes. Hence every y_{kl}-coefficient of (18) must be at least as large as the corresponding coefficient of (18a). That is, (18) uniformly dominates (18a) over the region of interest $(y \geqq 0)$; it is a "tighter" support for the function $T(\cdot)$. The more commodity classes there are the greater will be the improvement of (18) over (18a).

The result that (18) dominates (18a) implies that the representation using (3) is to be preferred over the "equivalent" more compact representation using (3a) and (3b). Any variant of Benders Decomposition should converge in fewer major

Table 4 First comparisonl of Benders decomposition for two alternative model representations

Major Iteration Number	Representation (3a) and (3b)		Representation (3)	
	LB	UB	LB	UB
1	—	5.410	—	5.053
2	4.150	5.023	5.000	5.028
3	4.349	"	≥5.008	(Convergence)
4	4.415	"		
5	4.534	"		
6	4.601	"		
7	4.631	"		
8	4.661	"		
9	4.714	"		
10	4.716	"		
11	4.750	"		
12	4.750	"		
13	4.774	"		
14	4.774	"		
15	4.808	"		
16	4.817	"		
17	4.817	"		
18	4.839	"		
	(No convergence)			

iterations for the first formulation than for the second. We have direct computational confirmation of this fact as a result of having turned to the first representation only after experiencing disappointing results with the second. Tables 4–6 show three approximately comparable disjoint regional optimizations using the original Benders Decomposition approach for both representations. We say "approximately" comparable because some internal parameters of the master problem algorithm were changed slightly during the time lapse between the runs, but we are confident that this does not alter the comparison significantly. The convergence parameter ε was set at 0.02 in all runs.

These comparative results indicate that the more compact representation consistently requires many more iterations for convergence, due principally to poorer lower bounds from the master problem. The time per iteration is approximately the same for both representations because the size and structure of the master problem and the individual transportation subproblems is exactly the same in both cases. Thus the representation using (3) is far superior. The other representation was all but unuseable in our application, considering the many validation and post-optimization runs required.

A closely analogous observation concerning the crucial importance of model representation has been reported recently by Beale and Tomlin [3]. They undertook to solve a practical problem concerning the optimal decentralization of office facilities using a direct branch-and-bound approach with a problem formulation which turns out to be very close to the one considered here. Their experience was that

Table 5 Second comparison of Benders decomposition for two alternative model representations

Major Iteration Number	Representation (3a) and (3b)		Representation (3)	
	LB	UB	LB	UB
1	—	5.134	—	5.083
2	3.892	"	4.937	4.960
3	4.245	"	≥4.940	(Convergence)
4	4.453	5.046		
5	4.534	"		
6	4.544	"		
7	4.574	5.043		
8	4.680	"		
9	4.680	"		
10	4.735	"		
11	4.735	"		
12	4.749	"		
13	4.749	5.027		
14	4.749	"		
15	4.759	"		
16	4.768	"		
17	4.768	"		
18	4.785	5.010		
19	4.785	"		
20	4.785	"		
	(No convergence)			

Table 6 Third comparison of Benders decomposition for two alternative model representations

Major Iteration Number	Representation (3a) and (3b)		Representation (3)	
	LB	UB	LB	UB
1	—	5.158	—	5.158
2	4.425	5.036	4.925	4.957
3	4.431	"	≥4.937	(Convergence)
4	4.436	"		
5	4.438	4.967		
6	4.461	"		
7	4.494	"		
8	4.494	"		
9	4.496	"		
10	4.505	"		
11	4.508	"		
12	4.512	"		
	(No convergence)			

the problem proved to be much more tractable computationally when some of their constraints like (3a) and (3b) were replaced by constraints like (3).[2]

[2] The authors are grateful to K. Spielberg for pointing out the following early reference containing related ideas: Guignard, M. and Spielberg, K. "Search Techniques with Adaptive Features for Certain Mixed Integer Programming Problems," Proceedings IFIPS Congress, Edinburgh, 1968.

In this connection, we would like to point out an interesting relation between the two representations which becomes pertinent when problems of this sort are addressed by LP-based branch-and-bound. It can be shown that the convex hull of the feasible solutions to (3a), (3b), (4), $x \geqq 0$ and $y = 0, 1$ is given by the constraints (3), (4), $x \geqq 0$ and $y \geqq 0$. Thus the common practice of dropping integrality requirements in order to produce an LP relaxation at each node yields a tighter relaxation when (3) is used than when (3a) and (3b) are used. The price of this tighter bound and the reduction in branching which it affords is, of course, the additional time required to solve a larger LP at each node. It seems probable that some mixture of the two representations will be superior to either one alone in terms of total computing time (e.g., the separability of (3), (3a) and (3b) with respect to l suggests that (3) might be used just for the l's corresponding to the largest total demand). This appeared to be the case in Beale and Tomlin's study. It should be emphasized that the extra size of (3) by comparison with (3a) and (3b) does not offer any difficulty whatever when Benders' approach is used, thanks to the analytic reduction which takes place prior to setting up the continuous subproblems to be solved at each major iteration. The ease with which Benders Decomposition can use such superior model representations is a comparative advantage over direct branch-and-bound which does not seem to be generally appreciated.

The theoretical result stated above also suggests a general methodology for discovering improved model representations: for various subsets of constraints involving some of the integer variables, try to explicitly derive the convex hull of the integer feasible points. Another related instance where this can be done is given in Geoffrion and McBride [13].

4.6 Conclusion

The major conclusion arising from this study is the remarkable effectiveness of Benders Decomposition as a computational strategy for static multicommodity intermediate location problems. The numerical experience quoted in §4.3 shows that only a few cuts are needed to find and verify a solution within one or two tenths of one percent of the global optimum. The same type of behavior was observed in another full-scale application carried out recently for a major manufacturer of hospital supplies with 5 commodity classes, 3 plants, 67 possible DC's and 127 customer zones. This behavior, together with the advantages of being able to decouple the multicommodity capacitated multiechelon transportation portion of the problem into a separate classical transportation problem for each commodity, yields an extraordinarily powerful computational approach.

The reasons why Benders' approach requires so few cuts for this problem class are not yet clearly understood. The discussion of §5 shows that one essential ingredient is making an appropriate choice among alternative mathematical representations of the same physical problem. We were able to employ a representation which incorporates the many constraints describing the convex hull of a portion of the

problem's integer feasible solutions. This was workable because of special opportunities for analytic simplification inherent in Benders' approach (it would not have been computationally feasible to use the same representation with a branch-and-bound approach to the problem). We hope that others will be motivated to study the questions raised by our observations with the objective of understanding more clearly the convergence behavior of Benders Decomposition and how to enhance it through appropriate choice of model representation.

Another conclusion we have reached on the basis of our experience is that every effort must be made to make it easy and economical to carry out the numerous pre- and postoptimality runs required to properly execute a practical application. This point, discussed in §4.2 and so well appreciated in the domain of linear programming, is rarely addressed in the existing integer programming literature. The burden of this requirement is exacerbated by the fact that many of the required runs must achieve very nearly optimal solutions if they are to be useful. This is certainly true of the probationary exercises, where significant suboptimality could shake management's confidence in the entire project, and is also true for "what if . . . ?," sensitivity, continuity, tradeoff and priority analysis runs as well because their very usefulness depends on the ability to measure *differences* between the solutions of different runs in a series. Obviously the tolerance on optimality must be quite tight if one is to avoid reaching spurious conclusions when making such comparisons. The results of §4.3 show that the approach developed here meets this requirement at reasonable computational cost.

The success with the present model suggests the desirability of expanding its scope. We shall mention here but two of the more appealing and easily accomplished possibilities. One is to include selection among alternative plant sites and plant capacity expansion projects via some additional 0–1 variables. Another is to take account of the service elasticity of demand, that is, of the fact that a customer zone's demand for various commodities tends to increase with the proximity of its assigned distribution center due to the advantages of decreased delivery delay [20], [22]. One way to incorporate this effect is to replace D_{il} in the model by D_{ikl}, the demand for product i by customer l if assigned to distribution center k. A (negative) net revenue term would also have to be appended to the objective function since total revenues to the firm would no longer be constant. Both of these extensions require but simple modifications to the algorithmic approach and do not upset the major factors controlling its efficiency (the use of a model representation yielding powerful Benders cuts and the separability of the multicommodity transshipment subproblem into an independent transportation problem for each commodity). We hope to be able to report on these and other extensions in a future paper.

References

1. Balinski ML, Spielberg K (1969) Methods for integer programming: Algebraic, combinatorial and enumerative. In: Aronofsky JS (ed) Progress in operations research, Vol. III, Wiley, New York

2. Bartakke MN, Bloomquist JV, Korah JK, Popino JP (1971) Optimization of a multi-national physical distribution system, Sperry Rand Corporation, Blue Bell, Pa. Presented at the 40th National ORSA Meeting, Anaheim, California, October
3. Beale EML, Tomlin JA (1972) An integer programming approach to a class of combinatorial problems. Math Programming 3(3)(December):339–344
4. Benders JF (1962) Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik 4:238–252
5. Davis PS, Ray TL (1969) A branch-bound algorithm for the capacitated facilities location problem. Naval Research Logistics Quarterly 16(3)(September):331–344
6. De Maio A, Roveda C (1971) An all zero-one algorithm for a certain class of transportation problems. Operations Research 19(6):(October):1406–1418
7. Ellwein LB (1970) Fixed charge location-allocation problems with capacity and configuration constraints. Ph.D. Dissertation, Dept. of Industrial Engineering, Stanford University, August
8. Ellwein LB, Gray P (1971) Solving fixed charge location-allocation problems with capacity and configuration constraints. AIIE Transactions III(4)(December):290–298
9. Elson DG (1972) Site location via mixed-integer programming. Operational Research Quarterly 23(1)(March):31–43
10. Fieldhouse M (1970) The depot location problem. University Computing Company, Ltd., London. Presented at the 17th International Conference of TIMS, London, July
11. Geoffrion AM (1973) Lagrangean relaxation and its uses in integer programming. Working Paper No. 195, Western Management Science Institute, UCLA, December 1972 (revised September 1973)
12. Geoffrion AM, Marsten RE (1972) Integer programming algorithms: A framework and state-of-the-art survey. Management Science 18(9)(May):465–491
13. Geoffrion AM, McBride RD (1973) The capacitated facility location problem with additional constraints. Working Paper, Western Management Science Institute, UCLA, December
14. Glover F, Karney D, Klingman D, Napier A (1974) A computational study on start procedures, basis change criteria, and solution algorithms for transportation problems. Management Science 20(5)
15. Graves GW, McBride RD (1973) The factorization approach to large-scale linear programming. Working Paper No. 208, Western Management Science Institute, UCLA, August
16. Gray P (1967) Mixed integer programming algorithms for site selection and other fixed charge problems having capacity constraints. Ph.D. Dissertation, Dept. of Operations Research, Stanford University, November 30
17. Khumawala B, Akinc V (1973) An efficient branch and bound algorithm for the capacitated warehouse location problem. Presented at the 43rd National ORSA Meeting, Milwaukee, May
18. Lea AC (1973) Location-allocation systems: An annotated bibliography. Discussion Paper No. 13, Dept. of Geography, University of Toronto, May
19. Marks DH, Liebman JC, Bellmore M (1970) Optimal location of intermediate facilities in a trans-shipment network. paper R-TP3.5 presented at the 37th National ORSA Meeting, Washington, DC, April
20. Mossman FH, Morton N (1965) Logistics of distribution systems. Allyn and Bacon, 245–256
21. Soland R (1973) Optimal facility location with concave costs. Research Report CS 126, Center for Cybernetic Studies, University of Texas at Austin, February
22. Willett RP, Stephenson PR (1969) Determinants of buyer response to physical distribution service. J Marketing Research VI(August):279–283
23. Williams AC (1973) Sensitivity to data in LP and MIP. Presented at VIII International Symposium on Mathematical Programming, Stanford, California, August

Chapter 5
Structured Modeling and Model Management

Daniel Dolk

Abstract We discuss Geoffrion's contribution to model management and the practice of modeling through his structured modeling formalism. We review the trajectory of structured model management research, enumerating the contributions and limitations of both structured modeling and model management in general. We summarize by suggesting how Geoffrion's work could be leveraged to contribute to a next generation of model management.

5.1 Introduction

It is a distinct pleasure and privilege to contribute to book honoring Art Geoffrion. My chapter discusses just one facet of the many areas where Art has made prolific research contributions, namely the foundations of modeling as embodied in his development of structured modeling. Structured modeling is essentially a formalism for meta-modeling which relies heavily upon the conceptual modeling practices used in information system design, especially those relevant to database design. I will provide a retrospective of structured modeling in the overall context of model management, which hopefully will serve as a comprehensible introduction to this research for those unfamiliar with it, highlight the substantial contributions Art has made in this field, and suggest ways in which structured modeling and model management may still be relevant today.

Modeling plays a central role not only in the disciplines of operations research and management science (OR/MS) but also in the process of information systems analysis and design. Indeed, modeling and simulation have become the third pillar of scientific inquiry in addition to theory and experimentation. Yet, though models are apparently sacrosanct in so many areas of intellectual endeavor, there appears to be no sense of urgency to cataloguing and managing the processes, contents, assumptions, results, and impacts of these artifacts we call models. This, despite the

Daniel Dolk
Department of Information Sciences, Naval Postgraduate School, Monterey, CA 93943, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_5,

significant body of work done in the area called model management in the last two decades of the 20th century.

The inability of model management to catch the attention of a broader community, particularly the organizations and associated decision makers who stand to benefit the most from it, and suffer the most from lack of it, is a curious phenomenon. One wonders whether the times have not yet caught up with this opportunity or whether there is a deeper cultural rift that leaves the art and practice of modeling beyond the pale of ordinary organizational concerns. Perhaps it is a propitious time to launch a modest retrospective of model management to ascertain what lessons may be learned and what promise, if any, the discipline may still hold. Specifically, I address the following questions: "Is model management still relevant?" "Can we reframe the basic objectives of this research to be relevant to contemporary network-driven, simulation-centric technologies?" "If so, what would it look like in today's landscape?"

In order to address these questions succinctly, Geoffrion's structured modeling [25] will serve as the operative lens. Although structured modeling is only one of many knowledge representation schemes for models, it is the most fully developed theoretically and practically, so it will serve by authorial fiat as the exemplar for model representation in this discussion. As a result, I will address the same questions specifically to structured modeling as to model management in the large.

5.2 A Brief History of Model Management

Around 1980, the research climate in the field of management information systems (MIS) was rife with opportunities. MIS itself was a brand new discipline, and there were cross-currents from many exciting developments taking place in other areas. Database theory as embodied in the relational data model [13] was still a very active arena (recall that the first commercial relational system did not appear until 1982). Artificial intelligence was experiencing a renaissance buoyed by a surge of optimism in the possibility of generating in silico human-like behavior. Decision support systems were also just emerging as a special class of information system transcending mere operational systems to provide more complex information to aid human decision making. The confluence of these streams of research in concert with the rapid development of computing languages and object-oriented methodology in computer science provided a wide open playing field for those who saw an opportunity for integrating information technology with existing OR/MS modeling techniques.

Model management was born from this landscape of developments in computer science and database management and was initially conceived as a modeling counterpart to data management [53]. The main tenet followed accordingly, namely that models, like data, should be treated as a shared corporate resource requiring systematic management and control. This would be aided and abetted by the functionality of a model management system (MMS) which would be the model counterpart of a database management system (DBMS). Implicit in this vision was the

recognition of the already existing rich vein of models and solvers which emanated from the OR/MS research and practitioner communities.

Buttressing this vision of model management was the concurrent emergence of the decision support phenomenon, which, in the spirit of [21, 48], posited models as the linchpins of decision making. Simon [49] in his seminal work portrayed a decision support system (DSS) as consisting of three major architectural components: data, models, and dialogue, each of which required an associated management system. Thus, the model management system was situated, conceptually at least, in a very strategic position as a promising research undertaking.

The corollary with data management naturally led to the question, "if we have a DBMS for the description, manipulation, and control of data, why not a model management system with the counterpart functionality for models?" As researchers began to think about what an MMS should be able to do, it quickly became clear that an MMS was a much more complicated beast than a DBMS. The prime directive for such a system was "support all phases of the modeling life cycle," which as Figure 5.1 shows entails significantly more than the data management dimension [41]:

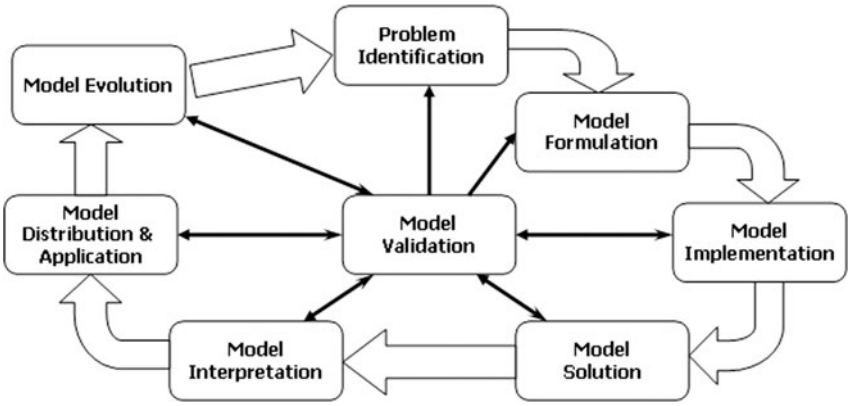


Fig. 5.1 Modeling life cycle (adapted from [41])

- *Problem identification* is similar to requirements specifications in information system development, wherein user/client requirements, model objectives, and data sources are identified.
- *Model creation* involves formulation of a conceptual representation of the model. Typically for OR/MS models, this representation consists of a mathematical description of the problem. However, as we argue below in the discussion of structured modeling, a conceptual model which subsumes the mathematical description as just one of many views of the overall representation is a highly desirable objective. Formulation may reuse an existing formulation, or incorporate a composition of two or more existing formulations, subject to revision and modification.

- *Model implementation* is the development of a computer executable representation of the model either through ad hoc program development or preferably using existing modeling languages and environments. Also, critically, this stage encompasses the identification, collection, and quality control of the associated data that will instantiate the model.
- *Model solution* requires identification of an appropriate solution algorithm, data preprocessing for providing input to the solver and delivering results to the database, and solver sequencing and execution.
- *Model interpretation* involves analyzing the results, understanding and debugging the model, and performing sensitivity analysis.
- *Model distribution and application* refers to the process of making a model operational and accessible to the user community on a need to know basis. Model and data security are mission-critical functions in the Internet age of information assurance.
- *Model evolution.* Model versions reflecting different sets of assumptions, data, and/or insights can proliferate rapidly and must be managed carefully. This may well result in a reformulation of the model, occasioning additional iterations through the life cycle process.
- *Model validation* is a persistent process occurring throughout the life cycle. This may range from dimensional and unit consistency analysis at the Formulation stage to the traditional internal and external validation processes at the Solution and Interpretation stages that the model solution is consistent with the initial assumptions and with the "real world."

In addition to the rather high-level life cycle requirements, more detailed functionality and design guidelines began to emerge as researchers delved more deeply into the architecture of an MMS. A majority of the work at this stage of research was focused upon environments for optimization models since they are generally well structured and there exists a large universe of models and solvers available for deployment. Some of the major requirements and associated guiding design principles emerged from the limitations of second-generation optimization software:

- *An MMS should have a uniform computer executable model representation.* The desideratum would be a representation formalism equal in power to the relational data model which underlies the database management environment. Additionally, the model representation should support multiple views of a model as the relational model does for data.
- *An MMS should support modeling languages.* The ability to describe models in a sufficiently general and abstract form, especially in a pseudo-mathematical representation, streamlines the ability to formulate models and widens the audience for model builders [22]. Earlier generations of software for optimization models, for example, required a matrix representation of models which one could correlate to programming assembly language in the software development arena. This restricted model formulation to a relatively small cadre of dedicated analysts.
- *An MMS should support cross-paradigm models.* The OR/MS world is an archipelago of modeling silos. A powerful alternative is a single environment which

simultaneously supports optimization, regressions, simulations, queuing, dynamic programming, etc., and reduces the need to learn a new software system for each different modeling paradigm. The potential of such a system to facilitate model integration is large. The OR/MS community has developed, and continues to develop, a broad portfolio of single paradigm models, access to any combination of which in a single environment would be a powerful tool in model integration (see below).

- *An MMS should have access to a library of solvers.* The OR/MS community has developed a multitude of solution algorithms and meta-heuristics for specific classes of models. As above, access to these solvers widens the range of models which can be usefully solved in this environment.
- *Models and model data should be separate in an MMS.* Early modeling systems required data to be in a very application-specific format which reduced or obviated the ability to reuse the data. Data should be independent of model representations until such time as a model instance is required. *An MMS should leverage relational technology for managing the data.* Data can be bound to a model representation using the powerful capabilities of relational databases. This underscores the separation of models and data mentioned above.
- *Models and solvers should be separate in an MMS.* Many models can be solved in different ways, for example, an LP model may be solved using simplex or branch-and-bound if an integer solution is desired. A model should not be bound to a particular solver until model solution time. The MMS should then be able to convert the data of the model instance to the appropriate format for the solver, and back again for the solution vector(s).
- *An MMS should support the reuse and integration of models.* Models are typically built for a single application and rarely ever reused beyond that application. The ability to reuse models not only has the potential for reducing model formulation costs but can significantly increase ROI from model development as well. Further, the ability to link existing models into composite models facilitates the development of more complex models.

The central theme which emerged from the list of requirements above was the need for a powerful model representation which is simultaneously comprehensible to a variety of different users (clients, analysts, mathematicians) and computer executable. The theoretical driver behind this quest started naturally enough with a database analogy: Is there a way to represent models that is comparable in power to the relational theory representation of data?

The first attempts at model representation leveraged artificial intelligence techniques for representing knowledge: semantic networks of nodes and edges for representing knowledge about models [19], first order predicate calculus to represent mathematical programming models in a way that allows useful inferences to be made [6] and frames for representing mathematical programming and econometric forecasting models [15]. Frames provided a basis for thinking about models in terms of object-oriented representations; many authors subsequently proposed various object-oriented representations for model management (see e.g., [24, 42]).

Early attempts to apply relational theory directly to model representation found that the transitive closure property which unifies relational theory does not carry over to its modeling counterpart [4, 5]. Thus, two or more models that are somehow joined with one another do not necessarily yield another model. The lack of a direct relational corollary led researchers to consider different alternatives.

Geoffrion developed a full model representation formalism called structured modeling based roughly on the entity-relationship data model, but which included significant extensions accommodating the ability to represent OR/MS models, particularly mathematical programming models [27]. Details of structured modeling will be discussed below, but it is interesting to note that structured modeling was the first contribution with respect to model representation which came from the OR/MS, rather than the information systems, research community. Other powerful representation techniques were developed as well including a relational version of structured modeling [16], logic modeling [2], graph grammars [37, 38], systems theory [45], and metagraphs [1].

5.3 Structured Modeling

Of the model representation approaches summarized above, Geoffrion's structured modeling has received the most attention by researchers. We provide a general recapitulation of this model representation formalism and show its vital role in the model management movement. This will by no means constitute a full and thorough review; readers are directed to [25, 27] for such a treatment.

Structured modeling (SM) is a semantic framework for representing wide classes of models, primarily from the domain of operations research and management science. Although many of the applications that structured modeling addresses in the research literature tend to be optimization models, Geoffrion went to great pains to show that models from a broad array of domains, some outside OR/MS altogether, could be represented using structured modeling.

SM has roots in the entity-relationship data model [12] but goes well beyond that in terms of formalism and extensions which accommodate modeling languages and indexing semantics. Every structured model is a collection of distinct elements, namely the *primitive entity* (/pe/), the *compound entity* (/ce/), and *attributes*. Attributes can be of four types: a regular *attribute* (/a/), a *variable attribute* (/va/) to designate decision variables in a model, a *function* element (/f/) based on the mathematical idea of a function, and a *test* element (/t/) which is a special Boolean case of a function, used, e.g., to represent constraints within optimization models.

Elements are grouped into classes called *genera*; a single such class is called a *genus*. A genus is in an "IS A" relationship with the elements comprising it. For example a supply constraint test genus may consist of an indexed set of supply constraint test elements, one for each supplier. Genera may be organized hierarchically to reflect high-level structures and to manage model complexity. This is done using

modules which are collections of genera constituting a subgraph of the overall genus graph (see below).

5.3.1 Structured Model Schema

Structured models are represented in three basic modalities: the schema, the genus graph, and the elemental detail (data) tables. The schema shown in Figure 5.2 is a full structured model schema representation of a simplified blending problem called FeedMix which determines amounts of materials to be used in animal feed that must satisfy certain nutritional requirements. The mathematical description of the model is

$$\min \sum_m Cm * Qm \tag{5.1}$$

$$\text{s.t.} \sum_m Aim * Qm \geq MRi \quad \forall i \tag{5.2}$$

$$Qm \geq 0 \quad \forall m \tag{5.3}$$

&FEED MIX MODEL

&NUTR_DATA NUTRIENT DATA
NUTRi /pe/ There is a list of NUTRIENTS.

MIN (NUTRi) /a/ : Real+ For each NUTRIENT there is a MINIMUM DAILY REQUIREMENT (units per day per animal).

&MATERIALS MATERIALS DATA
MATERIALm /pe/ There is a list of MATERIALS that can be used for feed.

UCOST (MATERIALm) /a/ : Real+ Each MATERIAL has a UNIT COST ($ per pound of material).

Q (MATERIALm) /a/ : Real+ The QUANTITY (pounds per day per animal) of each MATERIAL is to be chosen.

NUTR_MATERIAL (NUTRi, MATERIALm) /ce/ : There is a combination of NUTRIENTS and MATERIAL such that each MATERIAL has some NUTRIENTS and each NUTRIENT in one or more MATERIALS.
ANALYSIS (NUTR_MATERIALim) /a/ : Real+ For each NUTR_MATERIAL combination, there is an ANALYSIS (units of nutrient per pound of material).

NLEVEL (ANALYSISi., Q) /f/ ; @SUMm (ANALYSISim * Qm) Once the QUANTITIES are chosen, there is a NUTRITION LEVEL (units per day per animal) for each NUTRIEI from the ANALYSIS.
T:NLEVEL (NLEVELi, MINi) /t/ ; NLEVELi >= MINi For each NUTRIENT there is a NUTRITION TEST to determine whether the NUTRITION LEVEL is at least as large as the MINI REQUIREMENT.

TOTCOST (UCOST, Q) /f/ ; @SUMm (UCOSTm * Qm) There is a TOTAL COST (dollars per day per animal) associated with the chosen QUANTITIES.

Fig. 5.2 Structured modeling schema for FeedMix model

where m = material, i = nutrient, Cm = unit cost of material m, MRi = minimum requirement of nutrient i, Qm = quantity of material m, Aim = amount of nutrient i in material m.

The schema contains a full description of each of the genera and modules according to a pre-specified format. Genus information in a schema includes the genus name, associated index(es), any other genera it depends upon, genus type (/pe/, /ce/, /a/, etc.), domain (Real, Integer, etc.), and computable function (for test and function genera). Module information includes name and description. By convention modules are designated by a leading "&." Underlined and capitalized words in the descriptions are intended to be the main identifiers for the associated genus or module.

Structured models are built from the primitive entities outward. Primitive entities do not depend on any other genera so they form the foundation of the model. A typical building sequence is to identify each primitive entity and its associated attributes (NUTR/MIN and MATERIAL/{UCOST, Q}, for example), any compound entities and their associated attributes (NUTR_MATERIAL and ANALYSIS in this example), test genera (T:NLEVEL), function genera (TOTCOST and NLEVEL), and modules (&NUTR and &MATERIALS). Note that NLEVEL is used to compute the test genus T:NLEVEL and is more of an intermediary computation whereas TOTCOST is a terminal computation (leaf node) which may likely serve as an objective function for the model. Typically model building is more easily accomplished using the genus graph than working directly with the schema representation.

The schema is a textual description represented as a hierarchical structure in outline form. Each entry in a schema line is either a *genus* or a *module* (*prefaced by "&"*) name. Modules are aggregations of genera and/or modules which allow subsets of a model to be collected into a higher order structure. &NUTR_DATA, for example, captures the part of the model that deals solely with nutrients, specifically the genera NUTR and MIN. The genus is the basic component and may be of several different types as designated within the "//" separators: primitive entity (/pe/), compound entity (/ce/), attribute (/a/), function (/f/), and test (/t/). Primitive entities will usually have an associated index which is specified as part of the name, e.g., MATERIALm. For mathematical programming models, primitive entities correspond to index sets.

Each genus has a *calling sequence*, which may be null, specifying all genera the current genus may reference. The genus MIN, for example, references the primitive entity NUTR since it is an attribute of NUTR; therefore, NUTRi is contained within its calling sequence. Note the indexes are carried forward into the calling sequences; the indexing specification can become quite convoluted ([32] for more details). Primitive entities have no calling sequences; they are in effect a root of the genus graph tree (see below).

Attributes are equivalent to parameters in math programming models and thus have a data type; in our example, all attributes are in the set of positive real numbers. Test and function elements such as T:NLEVEL and TOTCOST, respectively, are typically described by equations represented in a modeling language (SML: Structured Modeling Language in this case).

Finally, each entry has a documentation segment with the potential for hyperlinking among them (words in all caps above). The documentation provides a medium for descriptions of the entry as well as any attendant model assumptions.

The schema is flexible in that different subsets, or views, can be displayed to appeal to different audiences. For example, one could provide an outline containing only the names and descriptions for end users and decision makers. Alternatively, one could suppress the documentation and only display the more analytical aspects of the model (material in bold face) for analysts and modelers.

5.3.2 Genus Graph

The genus graph shown in Figure 5.3 is an acyclic directed graph which shows the relationships among the various model genera as specified in the schema. It resembles an entity-relationship diagram in many respects. Entities come in two forms: primitive entities and compound entities. Note that the root nodes of the tree are the primitive entities which in the case of optimization models typically correspond to the indexes of the mathematical model. Compound entities represent the equivalent of relationships between two or more primitive and/or compound entities. In Figure 5.3, for example, the compound entity NUTR_MATERIAL represents two relationships between the primitive entities NUTR and MATERIAL, namely "each MATERIAL contains one or more NUTRIENTs" and "each NUTRIENT may be present in one or more MATERIALS." Genera which depend on other genera such as NUTR_MATERIAL are connected by a directed arc to the antecedent genera. Structured modeling does not provide an explicit way to designate the cardinality of

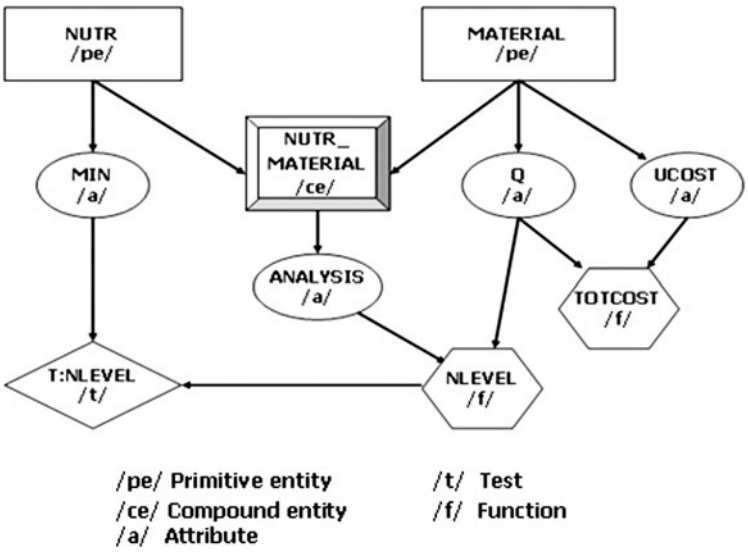


Fig. 5.3 Structured modeling genus graph for FeedMix model

a relationship between two entities as is the case with entity-relationship diagrams where each arc is labeled with the cardinality.

The parameters of the mathematical model are attributes of primitive or compound entities such as Q, the quantity of MATERIAL. Other dependencies cascade down the graph to the leaf nodes, which typically correspond to the constraints and objective function of the corresponding optimization model. We must emphasize, however, that the structured model does not specify the decision variables or objective functions explicitly. This binding takes place only at solution time when the user identifies the objective function(s), the constraint(s), the decision variable(s), and the solver, perhaps using a notional modeling language statement similar to

```
SOLVE FEEDMIX
MIN TOTCOST
SUBJ_TO T:NLEVEL
VARYING Q
USING CPLEX
```

5.3.3 Elemental Detail

The elemental detail aspect of a structured model is the relational table equivalent of the genus graph. Like entity-relationship diagrams and unified modeling language (UML) diagrams, a relational database schema of tables can be automatically created from a well-formed genus graph. The associated set of tables for the FeedMix genus graph is shown in Figure 5.4. These tables can be populated manually or by using more sophisticated SQL and XML commands to import data from external source databases. Note that this enforces the independence between model representations and data in the sense that we can add nutrients and/or materials by simply making additional entries in the data tables without changing the model representation at all.

A model that has its elemental detail tables populated is called a *model instance*. A model instance will be solved as indicated above, which in turn requires the ability to convert the elemental detail tables into the format required by the solver and then

NUTR Table

NUTR	INTERP	MIN	NLEVEL	T:NLEVEL
P	Protein	16	15.00	FALSE
C	Calcium	4	4.50	TRUE

NUTR_MATERIAL Table

NUTR	MATERIAL	ANALYSIS
P	std	4.00
P	std	14.00
C	add	2.00
C	add	1.00

MATERIAL Table

MATERIAL	INTERP	UCOST	Q
std	Standard	1.20	2.00
add	Additive	3.00	0.50

TOTCOST Table

TOTCOST
3.90

Fig. 5.4 Elemental detail (data) tables for the FeedMix model

conversely to translate the solution from the solver back into the elemental detail tables. In our example the column Q in table Q and the singleton table TOTCOST would both receive values from the solver indicating the optimum levels of Q and resultant TOTCOST, respectively.

5.3.4 Modules

Structured modeling also supports the hierarchical decomposition of models and the notion of multiple views via Modules which allow users to group related genera into a compressed subgraph for subsequent drill-down. Figure 5.5 shows the modularization of the Nutrient and Materials components of the FeedMix model.

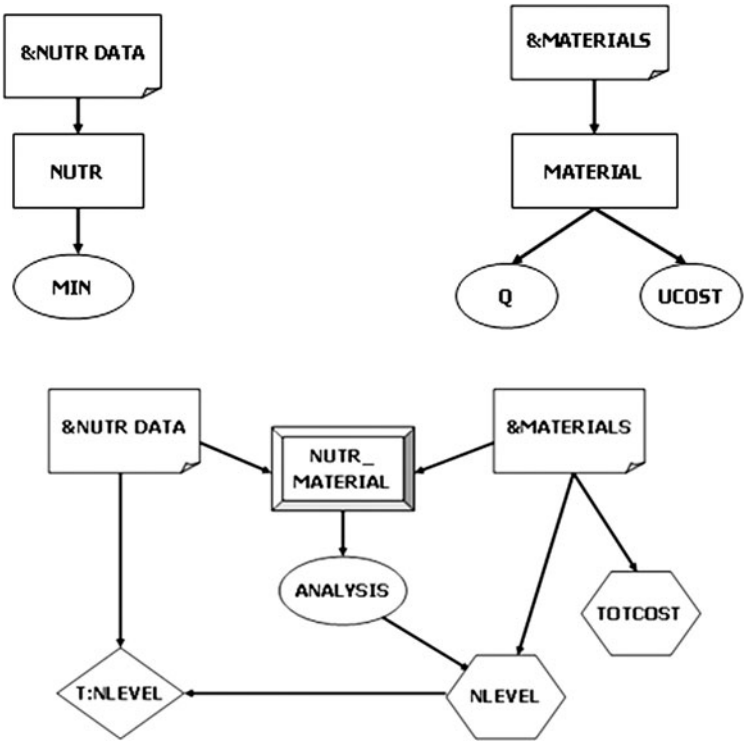


Fig. 5.5 The use of modules to compress the FeedMix model

5.3.5 Structured Modeling Language (SML)

Structured modeling strongly supports the technology of modeling languages through its Structured Modeling Language (SML) [30, 31]. The objective of algebraic modeling languages is to formulate models in relatively abstract,

quasi-mathematical form which allows a parsimonious, computer executable description of a model. The effect of such modeling languages is to place a higher level of the processing costs on the modeling software rather than the user. SML differs from two of the most popular algebraic languages, GAMS and AMPL, in that it includes four different semantic levels, of which only one is algebraic. Level 1 is for simple, definitional systems and directed graph models. Level 2 includes Level 1 plus the ability to express numeric formulae and propositional calculus expressions. Level 3 encompasses Level 2 with simple indexing capabilities as well as predicate calculus expressions. Level 4 subsumes Level 3 plus the ability to handle more complex indexing expressions as well as relational and semantic database models. SML levels are upward compatible in that a model expression at any level is valid at any higher level. The FeedMix schema shown above is an example of Level 3 SML.

5.3.6 Structured Modeling Environments

As research in structured modeling evolved, the concept of a model management system also evolved into a more general version called a modeling environment or integrated modeling environment [28]. A modeling environment is based less upon the notion of a single, stand-alone RDBMS-like counterpart and more upon the premise of a resource-rich infrastructure for supporting most, if not all, of the modeling life cycle as shown in Figure 5.1. Modeling environments may even be relatively application specific (e.g., [43]), but nevertheless transcend the narrow, single-platform focus of existing decision support and modeling software.

Many prototypes of structured modeling environments have been developed with varying degrees of success; we will mention only a few here (see [33] for references to some of the earlier versions). Geoffrion's own FW/SM [29] was built using the Framework system, a personal productivity tool from the late 1980s. Framework was an unfortunate choice for such a prototype because it lacked scalability and also had a short shelf-life in the software marketplace. Nevertheless, the platform was coaxed into supporting a robust version of structured modeling which did all the complex parsing of schemas and was able to generate database schemas automatically from the structured modeling descriptions. This proof of concept enforced many of the design principles mentioned in Section 5.2 such as model–data independence, model–solver independence, leveraging relational databases for data management, and linking with actual solvers in real time. There was not, however, a graphical user interface to ease model formulation, a serious shortcoming which we address in more detail later.

An ambitious implementation of structured modeling in the form of VMS/SM (Visual Modeling System for Structured Modeling) did include a GUI and implemented a substantial subset of structured modeling principles [55]. The GBMS/SM prototype also featured a genus graph GUI for the specification of structured models [11]. Later, a spreadsheet-based version of structured modeling was

developed as part of providing service-oriented, Web-based model management [36]. Their prototype GUI, although a simplification of the genus graph, is notable for using spreadsheets as the underlying platform for model management and structured modeling implementation. It would have been most interesting to see whether a GUI-driven FW/SM using a spreadsheet rather than Framework would have gained more traction for structured modeling and model management in general. Several other prototypes emerged based on object-oriented methodologies, including the ASUMMS/DAMS project [47] and BLOOMS [24]. All of these implementations represented valiant attempts at producing a generalized modeling environment for widespread usage; however, none of them survived with one notable exception.

The most enduring implementation of structured modeling is the Structured Modeling Technology (SMT) project at IIASA [43]. SMT supports a very large and complex optimization model called RAINS which is used to support international negotiations over European air quality. RAINS consists of several sub-models containing over 30,000 variables and 30,000 constraints in aggregate. Updated versions of the model will be even larger. Some of the interesting aspects of this implementation are as follows:

- No genus graph GUI has been implemented. The rationale for this design decision was that the complexity of the model in terms of the number of variables and their attendant interactions prevents a concise graphical representation. Even with larger contemporary monitors, one could see only an insignificantly small subset of the model at any one time.
- The centrality of data management. SMT leverages relational DBMS technology heavily to integrate and manage not only data sets but also model versions, model results, and model documentation.
- Multiple views of models and data. SMT supports a large community of users with diverse requirements. This results in the need for viewing models from a number of different perspectives. The structured modeling representation is leveraged to provide these model views.
- Open source. SMT subscribes to the open source philosophy to make the platforms as universally accessible as possible.
- No dimensional or unit specifications. An attempt at implementing units for each of the parameters led to the proliferation of such complex, non-intuitive unit specifications in the composite variables that this effort was eventually abandoned.
- Enforced documentation. SMT generates automatically human-readable documentation at each step of the model life cycle process.
- No model integration. The SMT philosophy is that it is easier to construct models from "scratch" than attempt to reuse already built models.

SMT remains heavily used to this day. It demonstrates the utility and scalability of the generalized structured modeling approach for large, complex optimization models with a diverse community of users. As a case study, it is invaluable in highlighting which of the tenets of model management and structured modeling are critical and which are optional or even dispensable.

5.4 Structured Modeling Contributions to Model Management

The contributions of structured modeling to model management are numerous. First and foremost, it provides a formal semantic ontology for models within a rigorously developed framework based on graph theory. In this formalism, mathematical models can be represented as conceptual models, thus unifying mathematical modeling as practiced in the OR/MS fields with the disciplines of information and data modeling. This cross-fertilization, it should be noted, was unusual in that it emanated from a research luminary in operations research adopting IS approaches rather than the more common situation at the time of IS researchers trying to extend data and information modeling concepts to operations research. Geoffrion's high standing in his field lent tangible momentum to the model management movement and generated a flurry of research on this topic from both communities.

Another advantage of structured modeling, as with most forms of conceptual modeling, is that it provides a bridge that allows analysts to communicate more effectively with decision makers. The ability to view the model structure as an influence diagram separate from the mathematical description provides, in principle, less analytically gifted players the ability to question assumptions and better comprehend the relevance of the model to the business environment. The commercial Analytica^{TM} system relies heavily upon this model representation which offers the advantage of a more "user friendly" view of mathematical models while simultaneously providing computer executability.

Another major contribution of structured modeling was the furthering of algebraic modeling languages as an accepted way of formulating computer executable mathematical models. The powerful modeling language, SML, with sophisticated indexing capabilities augmented the seminal work done in the development of the GAMS [9] and AMPL [23] languages. Although we take such modeling languages for granted today, the path from "horseblanket" matrix generators to algebraic representations was a rather slow and arduous process that spanned quite a few computing generations. The emphasis on model representation which structured modeling embodied served as an important catalyst for this transition.

Structured modeling also contributes to the goal of model reuse and integration. The lack of reuse is frequently cited as one of the factors in the relatively high cost of model development. In the same way that relational databases free data from being too tightly tied to specific applications, it is also desirable for models to escape the "one time only" application label. As shown in [26], genus graphs can be saved in model libraries and ultimately reused either by revising existing templates for applications with similar assumptions and structures or by composing more elaborate models from the linkage of two or more templates. Although genus graphs provide a fruitful medium for identifying potential sources of this latter form of model integration, it is by no means an easy task to carry out the integration itself. Not surprisingly, it appears this process cannot usually be done completely automatically but requires manual intervention to resolve semantic incongruencies between genus graphs, for example, in the resolution of naming differences and dimensional inconsistencies [7, 3].

There is a very large body of research dealing with structured modeling and related model management issues, much of it chronicled by Geoffrion [33]. I believe it is safe to say that this research has demonstrated structured modeling to be a powerful representational basis for comprehensive modeling environments that support OR/MS models. However, it is also safe to say that, with a few exceptions, structured modeling is not widely used today and never gained wide acceptance even among practitioners in the OR/MS community. The following section will attempt to address why this is the case.

5.5 Limitations of Structured Modeling

The limitations of structured modeling and the obstacles toward adopting it as a standard model representation form can be attributed to both endogenous and exogenous factors. Among the inherent limitations of structured modeling which have impeded its use by a wider audience is the high degree of complexity of the schema representation itself, particularly with respect to indexing semantics. Even seasoned modelers have been known to struggle when working at the structured modeling schema level. Generating a formal, correct schema even for a simple model such as that shown in Figure 5.2 is challenging in that each segment of the schema has its own precise syntax which leads to a steep learning curve.

One obvious way to mitigate this problem would have been to create a graphical user interface (GUI) for constructing genus graphs which could automatically generate most of the needed syntax. Although several graphical prototypes of structured modeling were implemented, e.g., [11, 36], none reached effective operational status. Surprisingly, Geoffrion's own FW/SM prototype did not include a GUI but rather required users to input the schema textually. As a result, there was never a full "cradle-to-grave" structured modeling implementation with a "user friendly front end." Without a GUI feature, structured modeling was regrettably confined to a relatively small cadre of experienced modelers and acolytes. It should be mentioned that at the height of the research into model management, graphics software was still a relatively immature technology. Graphics libraries were relatively expensive, typically not widely available or particularly easy to use, and tended to be confined to high end, programming intensive applications. There was no Visio™ equivalent at that time which could have solved this problem handily.

Another major shortcoming of structured modeling is that its strength lies in the representation of static models vis-à-vis dynamic models. It is not by chance that optimization models were the most successfully rendered examples in structured modeling. Their high degree of structure is well suited to the structured modeling formalism. However, when we consider the class of discrete event simulation models, for example, representation becomes inexorably more complex. Now we must deal not only with structures but also with events and processes that are time driven as well. There are no available means within structured modeling for representing event-driven processes in an elegant or concise fashion. Nor are there ways to incorporate the stochastic nature of such models. Although some suggestions were

made to address this shortcoming such as extending structured modeling to include a new random-valued attribute genus type [46], this avenue of research never gained a foothold.

In this vein, technology to some extent overcame structured modeling as well. During the past decade, the object-oriented Unified Modeling Language (UML) has become the prevalent data modeling methodology, overtaking the entity-relationship approach in system development. Although UML is anything but a polished model representation system, it does provide capabilities for representing both the static and dynamic dimensions of a model. Nevertheless, UML still does not offer a natural way to represent decision models in its environment [17].

There were also significant exogenous factors affecting the acceptance of structured modeling. Perhaps of primary importance from an organizational perspective is that most organizations do not support a modeling culture in which models are viewed as a sustainable asset. Models are too often consigned to spreadsheet exercises or ad hoc, application-specific projects. Thus, it is difficult in this setting to see any payoff for a general methodology such as structured modeling.

In the academic world, a cultural factor weighing against the adoption of structured modeling is that OR analysts, who would comprise the most likely user community, tend to be primarily mathematically trained and solver oriented. In that user group, mathematics is the lingua franca of model representation, and reframing the formulation of models into a conceptual modeling context is likely to be seen as an extra, undesirable layer in model development. Structured modeling forces general model structure to be specified before any model instance can be specified, and this is resisted for all the same reasons that documentation of models (and software) is always resisted. Even in the field of database design, the birthplace of conceptual modeling, database analysts often circumvent this phase and jump directly to building tables. Perhaps then, structured modeling might best be used in the classroom as a way of developing sound modeling practices that transcend simply the mathematical dimension.

5.6 Limitations of Model Management

One cannot easily view structured modeling outside the context of model management, and "model management" is a term that is rarely used today. It is neither commercially nor academically viable. Both model management and structured modeling faded away with the advent of the Internet and distributed computing. Some of the reasons for this are conjectured below.

- *No demand for MMS.* To the chagrin of those involved in this research over the years, it is not clear that there is, or ever was, a market demand for a model management system [54]. Although researchers heroically assumed that such an artifact would be valuable, no business value proposition was ever formulated or tested to verify this hypothesis. Few organizations support a modeling culture to the extent that development or purchase of a model management system could be

easily justified. Further, for many organizations, modeling begins and ends with the spreadsheet, so it would not be an exaggeration to say that, in the commercial world, "the spreadsheet is the MMS." Unfortunately, the spreadsheet is a fairly primitive modeling platform, which suffers from a spate of problems not the least of which is the widespread abuse of modeling practices by vast hordes of amateur modelers. Nevertheless, I believe the model management community could have benefitted, and could still benefit, from more research into the use of the spreadsheet as a *model management generator*. The spreadsheet-driven structured modeling prototype described in [36] not only showed the promise of this approach for smaller applications but also demonstrated that a conceptual modeling interface could serve as an effective vehicle for enforcing improved integrity in model formulation. With the advent of more scalable spreadsheets in contemporary software suites, much larger models would lend themselves to this approach as well.

- *Modeling too infrequently used in decision making.* This is the eternal complaint of the OR/MS community so shall not be dwelt upon further here. Despite the best efforts of our MBA and masters programs in management, end users of models, namely decision makers in organizations, are all too often either "model adverse," or less charitably, "model challenged."
- *Cross-paradigm myopia.* Selling model management in the academic community was not much more successful than in the commercial marketplace. Even in the analyst world where one might expect a more cordial reception to the concept of an MMS, cross-paradigm myopia tends to be prevalent, and this undermines the objectives of a generalized system. People tend to see the world in terms of the modeling paradigm in which they specialize, whether it be optimization, multi-criteria decision analysis, simulation, statistics, etc., and subsequently become familiar with one or two "stand-alone" software systems which solve only those kinds of models. The benefits of a generalized system which could handle models across multiple paradigms as is required for integrated supply chain management, for example, are thus not highly valued, and the stovepipe mentality persists.
- *Data but not models.* Models in general do not command the same respect as data in organizations. Everyone in this age is familiar with the need for data management, the value of institutional data for data mining purposes, and the challenge of and necessity for data security. A similar awareness about models is simply not prevalent; the fact is that the basic assumptions about model management articulated 30 years ago simply do not hold up in practice. Perhaps this will change as information technology evolves, but it is difficult to be confident about this conjecture. Underscoring this pessimism is the current landscape where enterprise resource planning (ERP) vendors are now including basic optimization models as part of their integrated architectures, particularly with respect to supply chain management. However, these models are essentially "black boxes" to the users who generally have little or no idea about the structure or details of the models being implemented and presumably used in decision making. In fact, this "stealth modeling" contravenes every principle of model management, obscuring rather than revealing the true purpose and value of modeling in an organization.

5.7 Trajectory of Model Management in the Internet Era

The concept of model management changed dramatically with the advent of the Internet. The emergence of the Internet shifted attention away from the generalized, monolithic system concept to a distributed resources perspective as we discuss below. Some of the well-known transitions which the Internet effected are shifts in perspective from stand-alone machine centric systems to network-centric systems, from top down to bottom up, from MMS as a single monolithic system to MMS as dynamic, configurable software components, from software as a product to software as a service, and from individual problem solving to collaborative problem solving.

In the model management domain this manifested in projects such as DecisionNet [3] which effectively decomposed model management systems into distributed resources managed by a centralized registry and directory. In this highly distributed environment, model representations, solvers, data, and sensitivity analysis software are all presumed to reside at distributed locations rather than in a centralized system. The environment is responsible for registering various resources, ensuring the availability of appropriate interfaces, and facilitating the necessary integration of resources for modelers to accomplish specific tasks in the life cycle. This network-centric version of model management views software as a service rather than a product, to be priced on a "per-use" basis as opposed to a one time purchase. In fact much of the work done in DecisionNet prefigured the more recent trends toward Web services and cloud computing.

The Internet also changed the kinds of models that organizations cared about. The flattening of organizational hierarchies manifested in Internet-based business led directly to a much higher requirement for, and interest in, collaboration, which, in turn, put a very strong emphasis upon business process and workflow models. These models are much more dynamic and therefore, akin to simulations than those with which structured modeling dealt.

The service-oriented Internet paradigm effectively eliminated most of the interest in the concept of a unified model management system. In principle, the ability to access modeling resources on demand, and *only* those resources which are required for any particular application at hand, is a much cleaner business model than the "all singing, all dancing" MMS. By the turn of the century, model management and structured modeling were no longer seen as central to the paradigm of dynamic, distributed computing. As a result, research interest in these areas per se began to fade and fragment into other related channels of inquiry. Nevertheless, pockets of research in model management persist to this day, and perhaps it is possible to build upon them and retrench as we discuss in the next section.

5.8 Next Generation Model Management

We now address the central question posed at the beginning, "Is model management relevant today?" Certainly research that incorporates this term has diminished

in recent years; however, we note that modeling itself is still a vibrant activity and continues unabated in organizations in the areas of information system development (e.g., business process modeling and enterprise integration architectures) as well as OR/MS-based decision models (e.g., supply chain management). The aggregate levels of modeling activity remain high, but the recognition of the need to manage these models still goes unheeded. We assume that model management is still a vital requirement but that we need to look at it in new ways that are consistent with advances in technology. We begin with the work that is still ongoing in model management augmented by suggestions which might be considered a partial blueprint for a "next generation model management".

5.8.1 Enterprise Model Management

The benefits of bringing model management to the field of enterprise and business process modeling are described in [34]. A unified enterprise modeling language (UEML, not to be confused with UML) is specified in [51] as a vehicle for bringing coherence to this endeavor in much the same way Geoffrion envisioned structured modeling serving the needs of the OR/MS modeling community. The UEML is intended to represent business logic in a platform-independent manner which nevertheless can be mapped to specific enterprise modeling toolkits and "that can, in theory, be merged, integrated, composed or otherwise operated upon to provide a larger subset of an enterprise model, thereby providing … a composed EM view of the enterprise" [34, p. 919]. This sounds very much like model–solver independence in the structured modeling world, model integration in the model management world, and service composition in the SOA world (see below). The UEML is itself a static conceptual model which could be rendered as a structured model schema thus possibly integrating decision models as an ingredient in the overall enterprise architecture. Regardless of the model representation employed, it is heartening to see that there is a realization of the need for model management in this arena. Hopefully, we can avoid reinventing the wheel and leverage the substantial model management research to move this agenda forward.

5.8.2 Service-Based Model Management

Another closely related opportunity for model management is the emergence of the service science, management and engineering (SSME) movement, which emphasizes service-dominant logic over the historically prevalent product-dominant logic [50]. This change in perspective from production to services changes the Producer–Customer relationship to a Provider–Consumer relationship in which both parties strive to "co-create value." This again will put emphasis on business process models, particularly collaborative models, but it will also require, in turn, a rethinking

of the more traditional quantitative modeling approaches which tend to optimize manufacturing efficiency over customer satisfaction.

On the information technology side of SSME, contemporary service-oriented architectures for delivering modeling and decision support will be required as well. Many of the problems faced by model management researchers in addressing the issue of model integration have resurfaced recently in the context of service-oriented architectures (SOAs). Specifically, an SOA must meet the challenge of composing services "on the fly" in order to satisfy a user's "on demand," and often ad hoc, request. This issue of service composition is very similar to that of model composition when trying to link existing models and data to satisfy a particular application (e.g., [10, 40, 45]). Interestingly, the SOA literature seems to show little, if any, awareness of the model management work already done in this area (e.g., [3, 35, 36]). However, research is beginning to resurface on service-oriented architectures for model management which redresses this situation [8, 14].

5.8.3 Leveraging XML and Data Warehouse/OLAP Technology

Even today, neither the entity-relationship model nor the UML model support the representation of decision models, so almost by definition there is room for a conceptual modeling approach which does. Because structured modeling has such a strong definitional character, it would also seem logical as a medium for some kind of ontological XML model interchange standard. This is consistent at a low level of implementation with the notion of model management as knowledge management proffered above. See [36] for an example of an XML representation of structured models.

Another immediate application for structured modeling would be to link SML representations with data warehouses and their associated online analytical processing (OLAP) environments. This provides an opportunity for accessing multidimensional data that align rather naturally with the mathematical index representation of OR/MS models [18]. Thus, it should be significantly easier to overlay a modeling system on a data warehouse than on a traditional relational database. The SMT system mentioned earlier adopts this approach, using a data warehouse as its data engine. In general, however, it seems that the modeling community has been slow to adopt this technology, and it is certainly the case that the data warehouse/OLAP vendors have been very slow to add significant modeling capabilities to their OLAP tools.

5.8.4 Model Management as Knowledge Management

As mentioned at the beginning of this retrospective, model management grew as a corollary to data management. Given the contemporary focus on knowledge flow,

the learning organization, and the knowledge society, a more robust metaphor may be model management as knowledge management. And even the term "management" has perhaps historically been used in too limited a scope, often implying a concern more with the management control issues of data, models, and knowledge than with their uses in conducting business more effectively.

When one looks at the modeling life cycle, it is difficult to escape the conclusion that modeling deals with anything less than the flow of knowledge. The purpose of building and solving models is to illuminate decision landscapes by identifying viable choices and evaluating trade-offs among the choice set. As such, models, properly fashioned, are knowledge creators. The stages of problem identification and model interpretation, for example, require extensive knowledge about the domain(s) for which the model is being developed. Model creation and solution, on the other hand, require specialized knowledge about mathematics and algorithms. Every aspect of the life cycle can be characterized in a similar knowledge-based idiom. I believe it is necessary to position modeling within the larger context of knowledge and to establish the "management" designation as the entire spectrum of management rather than focusing solely upon the control aspects. This perspective is examined in more detail in [44, 52].

Models from the decision support perspective are primarily used to guide and enlighten decision making. Too much of decision support, however, has dealt with isolated decision situations, for example, budget planning or optimal resource allocation. Experience with information system development over the past 15 years has shown us the primacy of business processes in the systems analysis process. Decisions need to be similarly cast not just as point events but as processes with overarching organizational objectives. Supply chain management is an excellent example often requiring multiple decisions which are highly interconnected and dependent on each other. The models which support these decisions must capture these interconnections and interdependencies. This puts a high premium upon model integration.

Simon's science of design philosophy has had a major impact on research in the areas of information system development, decision support, and operations research/management science. Too often, however, the focus from his approach has been on individual decision making in relatively narrow contexts. The Internet with its flattening of hierarchies increases the criticality of collaborative decision making, and the "new" model management must be able to marshal the flow and synthesis of modeling knowledge within collaborative environments.

At a higher level of knowledge sharing, one might imagine a Wikipedia counterpart for models, a knowledge-based equivalent of an open source operating system perhaps, where a community of committed scholars and practitioners creates and sustains an encyclopedia of information about a particular domain of applications and their attendant models. Consider supply chain management as an example. A knowledge environment for the supply chain world might contain a high level concept map, a semantic network of sorts, containing among other things, a taxonomy of supply chain functions. The concept map would contain for each function links to case studies, scholarly articles, and models suitably represented in structured

modeling or an equivalent robust representation. These models could serve pedagogical purposes as well as be computer executable in the distributed computing sense of DecisionNet.

5.8.5 Search-Based Model Management

Although decision support speaks of semi-structured and unstructured decision making, the majority of applications developed deal with quite structured data. The emergence of search engine technology provides a powerful capability to manipulate semi-structured data, especially documents. Thus we need to reconsider what models might look like that use these semi-structured data rather than, or in addition to, the typical data stored in relational databases. Can this kind of knowledge be used to refine model assumptions, amplify model interpretation, guide us to new solution heuristics, or build model taxonomies for model formulation? The "new" model management must leverage search engine technology to access and manipulate a wider range of knowledge. One interesting avenue of attack in this regard is the use of "mash ups" for identifying domain-specific indicators which could be marshaled toward a preliminary stage of model formulation; [39] gives an example of this in the area of clean energy.

5.8.6 Computational Model Management

For structured modeling to be truly relevant, I believe it is necessary to revisit the dynamic model representation issue to examine whether structured modeling can be extended elegantly and naturally to accommodate dynamic models such as discrete event and agent-based simulation models. The prevalence of computational modeling, particularly in the biological sciences, with its emphasis on bottom up complexity, cellular automata, agents, and emergent behavior, presents a distinct challenge to the relatively static forms of structured modeling. Yet this form of modeling seems to be gaining ascendancy and may lead us into another complete iteration of how we view models, what model management entails, and what modeling environments will look like.

In the new perspective on complexity that has resulted from research in the evolutionary and biological sciences, systems are simulated as "bottom up" phenomena, often represented as cellular automata, exhibiting emergent macro behavior from the repeated interaction of localized agents following (usually) relatively simple rules [20]. Interestingly, however, the software platforms which support these classes of agent-based simulations (ABS) seem hauntingly reminiscent of optimization software as it existed before modeling languages or representational formalisms such as structured modeling were developed. Other than the use of object-oriented architectures, there do not seem to be any uniform model representations for agent-based

models and simulations, and each model is built in an ad hoc, stand-alone mode. Each platform has its own protocol for representing and constructing simulations, and oftentimes its own community of practice for sharing knowledge. Given the extensive interest that currently exists about ABS, perhaps there is an opportunity for applying model management design principles that can accelerate the evolution of this modeling paradigm.

Agent-based simulations have evolved from cellular automata into elaborate virtual environments which pose different challenges for model management, especially around the issue of external model validation. This proliferation of what is sometimes termed computational modeling quickly outstrips the older, rather static, notions of model management and requires a more fluid, knowledge-based approach for the related processes of computational experimentation and computational explanation. The "new" model management must be able to handle a much more dynamic kind of model and oftentimes a fuzzier notion of validation, while perhaps simultaneously relying upon the more conventional OR/MS models as validation and calibration vehicles.

5.8.7 Model Management: Dinosaur or Leading Edge?

It is difficult to know, even in retrospect, whether model management has been overcome by events or whether it perhaps may still be ahead of its time. One can see in the agent-based simulation environment and the enterprise modeling endeavors the same phenomenon which occurred in the evolution of OR/MS modeling languages. The unregulated proliferation of different low level languages and methodologies, each with its relative strengths and weaknesses, leads to a recognition of the need for some uniform, integrative, higher level modeling methodology ("meta-models" in today's terminology) which allows a wide range of models to be described in "business friendly" or "decision friendly" terms while simultaneously being computer executable. Structured modeling, among other methodologies, played this evolutionary role in the world of OR/MS models, and it will be interesting to see whether it, or derivatives thereof, may eventually find purchase in other environments.

5.9 Summary

Geoffrion's foray into meta-modeling via structured modeling represented a significant intellectual departure for the majority of the operations research community which, at least in the mathematical programming arena, typically focused principally upon generating and demonstrating the relative efficiency of new solution algorithms and meta-heuristics. This shift of attention from model solution to the overall modeling life cycle and subsequently to the conceptualization of models and modeling languages was strongly cross-fertilized by the disciplines of

computer science and information systems. Structured modeling is an admirable blend of operations research, management science, programming languages, database management systems, software engineering, and information systems modeling. This union, it seems to me, has been underutilized by all the communities involved. The processes of building models and building information systems are remarkably similar as are, in a more focused context, the processes of building solvers and writing application programs. Yet, too often the software engineering world all but ignores model-based decision making, and too often the OR/MS world ignores proven system and software development methodologies in the course of model building. Geoffrion's signature contribution in the creation of structured modeling was to illuminate both these landscapes and show where and how they could be fruitfully combined.

I would like to conclude on a personal note of deep gratitude. My own career would not have been nearly as enjoyable or as fruitful without Art Geoffrion's friendship; working with Art and the structured modeling community he generated has been the most rewarding part of my academic life. It has been a distinct honor to see a Master Scholar at work and to be invited to participate in some of that work. I celebrate Art for his mentorship and support, for his boundless intellectual energy, his ability to see beyond boundaries and across the horizon, and for being a generous, wise, congenial, and committed colleague.

References

1. Basu A, Blanning R (1994) Model integration using metagraphs. Information Systems Research 5(3):195–218
2. Bhargava H, Kimbrough S (1993) Model management: An embedded languages approach. Decision Support Systems 10:277–299
3. Bhargava H, Krishnan R, Muller R (1997) Decision support on demand: Emerging electronic markets for decision technologies. Decision Support Systems 19:193–214
4. Blanning R (1982) A relational framework for model management. DSS-82 Transactions: 1:16–28
5. Blanning R (1985) A relational framework for join implementation in model management systems. Decision Support Systems 1:69–81
6. Bonczek R, Holsapple C, Whinston A (1978) Mathematical programming within the context of a generalized data base management system. R.A.I.R.O. Recherche Operationelle/Operations Research 12(2)(May):117–139
7. Bradley G, Clemence R (1987) A type calculus for executable modeling languages. IMA Journal of Mathematics in Management 1(4):277–291
8. Brodsky A, Al-Nory M, Nash H (2008) Service composition language to unify simulation and optimization of supply chains. Proceedings of the 41st Annual Hawaii International Conference on System Sciences, Hawaii, January 2008
9. Brooke A, Kendrick D, Meeraus A (1992) GAMS: A user's guide. Release 2.25. The Scientific Press, San Francisco, CA
10. Chari K (2003) Model composition in a distributed environment. Decision Support Systems 35:399–413
11. Chari K, Sen T (1998) An implementation of a graph-based modeling system for structured modeling (GBMS/SM). Decision Support Systems 22(2):103–120

12. Chen P (1976) The entity relationship model: Toward a unified view of data. ACM Transactions Database Systems 1(1):9–36
13. Codd E (1970) A relational model of data for large shared data banks. Communications ACM 13(6)(June):377–387
14. Deokar A, El-Gayar O (2008) A semantic web services-based architecture for model management systems. Proceedings of the 41st Annual Hawaii International Conference on System Sciences, Hawaii January, 95
15. Dolk D, Konsynski B (1984) Knowledge representation for model management systems. IEEE Transactions on Software Engineering SE-10(6):619–628
16. Dolk D (1988) Model management and structured modeling: The role of an information resource dictionary system. Communications ACM 31(6):704–718
17. Dolk D, Ackroyd M (1995) Enterprise modeling and object technology. Proceedings of the 3rd International Conference on Decision Support Systems, Vol. 1, Jun 22–23. Elsevier, Hong Kong, pp. 235–246
18. Dolk D (2000) Model integration in the data warehouse era. European Journal of Operational Research 122(April):199–218
19. Elam J (1980) Model management systems: A framework for development. Proceedings of 1980 Southwest AIDS Conference, Atlanta, GA
20. Epstein J (2007) Generative social science: Studies in agent-based computational modeling. Princeton University Press, Princeton, NJ
21. Forrester J (1958) Industrial dynamics: A major breakthrough for decision makers. Harvard Business Review 36(4):37–66
22. Fourer R (1983) Modeling languages versus matrix generators for linear programming. ACM Transactions on Mathematical Software 143–183
23. Fourer R, Gay D, Kernighan B (1993) AMPL, A modeling language for mathematical programming. The Scientific Press, San Francisco, CA
24. Gagliardi M, Spera C (1997) BLOOMS: A prototype modeling language with object oriented features. Decision Support Systems 19(1):1–21
25. Geoffrion A (1987) An introduction to structured modeling. Management Science 33(5)(May):547–588
26. Geoffrion A (1989) Reusing structured models via model integration. Proceedings of the 22 Annual Hawaii International Conference on System Sciences. IEEE Computer Society Press, Los Alamitos, CA, pp. 601–611
27. Geoffrion A (1989) The formal aspects of structured modeling. Operations Research 37(1)(January–February):30–51
28. Geoffrion A (1989) Computer-based modeling environments. European Journal of Operational Research 41(1)(July):33–43
29. Geoffrion A (1991) FW/SM: A prototype structured modeling environment. Management Science 37(12)(December):1513–1538
30. Geoffrion A (1992) The SML language for structured modeling: Levels 1 and 2. Operations Research 40(1)(January–February):38–57
31. Geoffrion A (1992) The SML language for structured modeling: Levels 3 and 4. Operations Research 40(1)(January–February):58–75
32. Geoffrion A (1992) Indexing in modeling languages for mathematical programming. Management Science 38(3)(March):325–344
33. Geoffrion A (1999) Structured modeling: Survey and future directions. INFORMS Interactive Transactions of ORMS. http://www.anderson.ucla.edu/faculty/art.geoffrion/home/biblio/text.htm, June
34. Goul M, Corral K (2007) Enterprise model management and next generation decision support. Decision Support Systems 43:915–932
35. Güntzer U, Müller R, Müller S, Schimkat R-D (2007) Retrieval for decision support resources by structured models. Decision Support Systems 43:1117–1132
36. Iyer B, Shankaranarayanan G, Lenard M (2005). Model management decision environment: A Web service prototype for spreadsheet models. Decision Support Systems 40:283–304

37. Jones CV (1990) An introduction to graph based modeling systems, Part I: Overview. ORSA Journal of Computing 2(2):136–151
38. Jones CV (1991) An introduction to graph based modeling systems, Part II: Graph grammars and the implementation. ORSA Journal of Computing 3(3):180–206
39. Kimbrough S, Lee T, Oktem U (2008) On deriving indicators from text. Wharton working paper, University of Pennsylvania, Philadelphia, PA, June 2008
40. Kottemann J, Dolk D (1992) Model integration and modeling languages: A process perspective. Information Systems Research 3(1)(March):1–16
41. Krishnan R, Chari K (2000) Model management: Survey, future directions and a bibliography. Interactive Transactions of ORMS 3(1):1–19
42. Lenard M (1993) An object-oriented approach to model management. Decision Support Systems 9(1)(January):67–73
43. Makowski M (2005) A structured modeling technology. European Journal of Operational Research 166(3)(2005):615–648
44. Makowski M, Wierzbicki A (2003) Modeling knowledge: Model-based decision support and soft computations. In: Yu X, Kacprzyk J (eds) Applied decision support with soft computing, vol. 124 of Series: Studies in Fuzziness and Soft Computing. Springer, New York, NY, pp. 3–60
45. Muhanna W, Pick R (1994) Meta-modeling concepts and tools for model management: A systems approach. Management Science 40(9):1093–1123
46. Pollatschek M (1995) SML for simulation. Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, 27 p
47. Ramirez R (1995) A management system for MS/OR models. Journal of Microcomputer Applications 14(2):53–60
48. Sprague R, Carlson E (1982) Building effective decision support systems. Prentice-Hall, Englewood Cliffs, NJ
49. Simon H (1969) The sciences of the artificial. MIT Press, Cambridge, MA
50. Vargo S, Lusch R (2004) Evolving to a new dominant logic for Marketing. Journal of Marketing 68:1–17
51. Vernadat F (2002) UEML: Towards a unified enterprise modeling language. International Journal of Production Research 40(17):4309–4321
52. Wierzbicki A, Makowski M (2000) Modeling for knowledge exchange: Global aspects of software for science and mathematics. In: Wouters P, Schroder P (eds) Access to publicly financed research. NIWI, Amsterdam, The Netherlands, pp. 123–140
53. Will H (1975) Model management systems. In: Grochia E, Szyperski N (eds) Information systems and organization structure. Walter de Gruyter, Berlin, pp. 468–482
54. Wright G, Chaturvedi AR, Mookerjee R, Garrod S (1998) Integrated modeling environments in organizations: An empirical study. Information Systems Research 9(1):64–85
55. Yeo G, Jian H (1997) Visual modeling with VMS/SM. Proceedings of the IASTED International Conference on Simulation and Modelling, Pittsburgh, PA, pp. 202–205

Chapter 6
Retrospective: 25 Years Applying Management Science to Logistics

Richard Powers

Abstract A management science practitioner recounts his 25 years of providing the corporate world with logistics optimization software and consulting. Clients included a substantial portion of the world's largest businesses as well as the US Department of Defense and General Services Administration. Significant contributions were made to the profitability and return on assets of these client organizations. At the same time the members of the author's company contributed to the ongoing development of optimization technology and large-scale data management to support logistics modeling. These efforts led to the publication of dozens of articles in first-rate logistics and management science journals as well as the election of two of the company's principals to the National Academy of Engineering.

6.1 Where It All Began

In the spring of 1975 I was working in the Office of the Secretary of Defense (OSD) when I received orders from the Chief of Naval Personnel to "report immediately" to a special study group that was being formed to revamp the logistics infrastructure of the Department of Defense (DOD). For the previous few years I had been working on the realignment of force structures of the military services as we departed from Vietnam and on the introduction of the All Volunteer Force (AVF). This new assignment would in a way be an extension of that work because the logistics infrastructure which was in place to support the war in Southeast Asia was still in place in 1975 although the force levels had been reduced significantly.

When I reported to the Department of Defense Material Distribution System (DODMDS) study, a group made up of about 50 military and civilian personnel from all of the military services and the Defense Logistics Agency (DLA), I was assigned the tasks of developing and applying analytical tools, defining and acquiring the data necessary to do the analysis, and managing the contracts that we would let with the private sector to assist our efforts.

Richard Powers
Former CEO and President, Insights, Inc., Redwood City, CA, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_6,

This admittedly seemed an overwhelming task to me at first blush. I had been doing manpower and cost modeling in OSD, but I had no relevant experience in the sort of resource allocation optimization implied by a "restructuring" of the DOD logistics system. In 1975 there were 34 wholesale distribution facilities across the four military services and DLA. Those facilities were scattered across the continental United States (CONUS) and Hawaii, with major concentrations along the coastal areas of the country. In 1975 there were 50,000 separate customers of the wholesale logistics system who received 27.4 million shipments worth almost $100 billion in 2008 dollars. Material moved into the DODMDS from 19,000 separate procurement sources. There were 3.7 million stock keeping units (SKUs) stored in 866 separate buildings within the 34 facilities. Just the base year warehousing and transportation costs, excluding inventory holding costs, were $4.6 billion in 2008 dollars.

As usual in Washington when the word gets around that a major effort like the DODMDS study is cranking up, numerous government contractors start pleading their cases about how they are the right ones to undertake this massive effort. We talked with all of those who claimed they knew just what to do, given that they received seven or eight figure contracts, but I was not convinced any of those contractors truly understood the magnitude of the task or had the tools and expertise to do it right. From day one we knew that it was highly likely that we had excess capacity in the DODMDS and that the results of our study would be the recommended closure of some of those DODMDS facilities, thus a political hot potato. This expectation pretty well guaranteed that our conclusions and recommendations would be severely scrutinized, challenged, and opposed as our report worked its way through DOD, OMB, and the Congress. For that reason I believed we should first explore conducting the data development and modeling work ourselves rather than just turn it over to a third party.

Fortunately there was a young Air Force officer, Lt. Jeffrey Karrenbauer, serving at the Air Force Logistics Command (AFLC) in Dayton, OH, who had recently completed his course work for a doctorate in logistics at Ohio State University, at that time arguably the best logistics academic program in the country. I requested that Karrenbauer be transferred to the DODMDS study group as he seemed to know a good deal about logistics modeling and the sorts of tools that were available to do it. Karrenbauer educated me about both goal-seeking and simulation models for logistics analysis and was aware of a recent article published in *Management Science* by Arthur Geoffrion and Glenn Graves at UCLA that appeared to hold some potential for our modeling requirements. We contacted Geoffrion and asked him to visit us in Washington to see if the approach he and Graves had developed could work for us. We concluded that adopting a location optimization model of the sort Geoffrion described was the way to go and set on that course and contracted with Geoffrion and Graves to work with us to do that. Looking inside the DOD for some further optimization expertise, we solicited the assistance of a relatively new member of the faculty of the Naval Postgraduate School in Monterey, Jerry Brown. Brown was a recent graduate of the doctoral program at UCLA and had worked with Graves and Geoffrion during his time there.

At the same time, we knew that to withstand the scrutiny and challenges that would inevitably come with results based solely on least cost we would have to be able to show that the structure we would recommend could support mobilization requirements in the time specified by the Joint Chiefs of Staff (JCS). To satisfy the questions about mobilization and operational requirements of our least cost structure we adopted a dynamic simulation model, LREPS, which had been developed at Michigan State University and was offered commercially by Systems Research in East Lansing. Meantime we had initiated a data call to all of the military services and DLA to provide to us all of the logistics transaction data for a year.

The DODMDS study group concluded its analytical work in the spring of 1978. We had processed over 3000 magnetic tapes of logistics data and consumed thousands of hours of large-scale computer resources at two DOD computer facilities. The two models we used served us well and gave us great confidence that we had done it right. Although our recommendations took several months to work their way through the various echelons of DOD, OMB, and Congress, the results were never successfully challenged on technical grounds. However, the political process in Washington has a way of altering and delaying the actions that appear to be warranted from a study as thorough and comprehensive as the DODMDS study. Nonetheless, over the roughly 20 years following the completion of the DODMDS study in 1978, virtually all of our recommendations were implemented in one form or another. I have no way of knowing the actual savings that did accrue for DOD and the American taxpayers over those 20 some years, but in 1975 we estimated that annual savings of 10% could be achieved by implementing our recommendations. The number of distribution facilities could be reduced by one-third and annual savings would be $500 million in 2008 dollars.

In the summer of 1978 I had 20 years service in the Navy. Those 20 years had been most rewarding and enjoyable, but as I thought about what we had accomplished with the DODMDS project I believed we should take the technology we had developed and the experience we had and take it to corporate America. Looking down the road it was clear that global competition was going to play a larger and larger role in the affairs of American companies. It seemed to me that if we could increase the productivity of American businesses, they could compete more effectively in this emerging global economy. So I decided to exchange my Navy blue uniform for pin stripes. Brown, Geoffrion, Graves, and Karrenbauer joined with me in 1978 to form INSIGHT, Inc., a company to be devoted to providing the best possible optimization-based management support systems to corporate America.

Within a couple of years we added several notable management scientists to INSIGHT'S professional stable of optimization expertise: Gordon Bradley and Rick Rosenthal from the Naval Postgraduate School, David Ronen from the University of Missouri, St. Louis, Richard McBride from USC, Shao Ju Lee from Cal State Northridge, John Mamer from UCLA, and Terry Harrison from Penn State.

Although we perceived some of them, but not all, a fortuitous confluence of factors was coming together in the late 1970s and early 1980s as we were getting started: a recognition of logistics as a crucial corporate function; the rise of finance as a driving force in corporate America; the emerging globalization of markets and

manufacturing; the spectacular increases in computing power; and the development of powerful new mathematical techniques for solving large, complex optimization problems of the sort encountered in logistics.

Before delving a bit into each of those converging forces let me explain my preference for the term "logistics network" rather than the frequently used term "supply chain." To me supply chain implies a hierarchical singularity that is seldom found in the business world, whereas a logistics network conveys the correct image of a highly complex, inter-related set of relationships within and between echelons of a sourcing, manufacturing, and distributing network.

6.2 The Rise of Logistics

Perhaps Peter Drucker was the first to see it in 1962 when he wrote about physical distribution as the "dark continent" of the US economy. The traditional business functions of warehousing and transportation were relegated to the bottom of the organizational hierarchy in most businesses. However, by the late 1970s and into the 1980s old hierarchical organization models for control and communications were proving to be inadequate in the rapidly changing global market place. Speed and flexibility became the name of the game, and the technologies of communications and computers were the enablers of that speed and flexibility. The old organizational structure of the pyramid, which had served well for decades, was too slow and cumbersome. Information systems replaced the middle echelons of the old pyramid structure, and the old ways of ordering business and making decisions about sourcing, manufacturing, and distribution were giving way to a need for integrating a diversity of players in a logistics network. Outsourcing many activities became an effective strategy and that required information and coordination. The logistics organization became that coordinating center for sourcing, manufacturing, and distribution. Wal-Mart recognized that before most other organizations. Wal-Mart recognized early-on the name of the game was logistics. Worth noting, Lee Scott, the third CEO of Wal-Mart following Sam Walton and David Glass, began with the company in the transportation department and moved up in the company through logistics.

6.3 The Rise of Finance

The early part of the twentieth century saw manufacturing people rising to the top of the corporate hierarchy—that was where the problems and leverage were as we became a mass-producing economy. After WWII manufacturing gave way to the rise of marketers to the top who could promote and sell the mass-produced goods. By the 1980s those with finance backgrounds were emerging to the top of many companies. With the emphasis on profitability and return on investment that came with greater numbers of powerful institutional investors, companies that wanted to

grow had to show returns on assets that would attract the capital they needed. As the well-known DuPont model made quite clear in simple terms, return on assets (ROA) is the product of capital turnover X profit margin. By finding effective ways to reduce capital committed to the business (close unneeded facilities, contract out warehousing, vehicle fleets, etc.) and to increase profit margins by decreasing all the operating costs associated with sourcing, manufacturing, and distributing products, ROA could be improved.

6.4 Globalization

The globalization of capital flows, manufacturing, and markets has had a profound effect on business organizations throughout the world. As Asian, Latin American, and European countries took on more and more of the manufacturing of products and components sold in the United States, the complexities of effectively managing these world-wide logistics networks became highly challenging. The amount of information required to envision the entire logistics network of a sizeable business was not only difficult to collect, it was equally difficult to assimilate in a meaningful way to support resource allocation decisions.

6.5 Computer Technology

Little needs to be said here other than that the incredible advances in computational capability over the past 30 years have been an absolutely critical enabler of the application of management science methods to complex logistics network issues. I do not think many people in 1978 really grasped the profound implications of "Moore's Law" for the kinds of complex logistics analyses that would be possible within a few short years. Today a reasonably well-configured laptop has more computing power than a roomful of mainframe computers in 1980.

6.6 Optimizing Solver Technology

Great strides were made in the late 1970s and early 1980s in creating optimizing solver technologies that enabled the solution of the huge and extremely complex resource allocation problems inherent in logistics networks. Mention was made earlier of the work of Geoffrion and Graves that was applied in the DODMDS project and that was followed quickly by network optimization codes developed by Brown, Bradley, and McBride. Of course, the rapidly evolving computer power was a potent enabler of these new optimizing solvers, which have continued a steady evolution ever since.

6.7 Insight Takes Off

With our objective of applying the best optimization technology available to the resource allocation problems of corporate America we launched in the summer of 1978. We were convinced that logistics management and management science were made for each other. INSIGHT's vision was from the beginning to marry research in large-scale optimization with application to real-world business problems. The manifestation of that vision and the delivery of INSIGHT's technology and expertise have changed with the times over the past 30 years, but the vision has been constant. INSIGHT has always had close ties to academia to assure the continuing focus on research. Although INSIGHT grew and added marketing, administrative and support staff, the heavy emphasis on research in large-scale optimization was maintained.

In the first few years, INSIGHT performed consulting engagements for corporate clients using our proprietary network optimization and data management software. Our earliest clients were Becton-Dickinson, Maryland Cup, and Glidden. The only computers capable of handling the huge databases and optimization programs were large mainframes. Our early computation work was done at UCLA and Geico Insurance where we bought computer time. When a logistics network optimization project started we would meet with the client management to get a clear understanding of their objectives and then work with them to build the model of their production–distribution system. We would then tell the clients what data they would need to provide us and set up a task plan to get the project done. After all the required data were collected and validated we would make a baseline model run, followed by optimization scenarios at our contract computer center. With printed outputs in hand we would then sit down again with client management to analyze modeling results and work through the "why" of those results. If more optimization model runs were necessary at this point we would do those and wrap up the project with a written report describing what had been done, the results, and recommended courses of action.

In 1980 Baxter Healthcare came to INSIGHT with a request to license our proprietary logistics network optimization software, ODS, and the data management software that built the databases and input files for ODS, DATA-1. (In 1984 ODS and DATA-1 were incorporated with a transportation simulator, SHIPCONS, into a fully integrated logistics network optimization and data management package, SAILS.) After creating some documentation and "hardening" the software, a license agreement was set up and Baxter became INSIGHT's first software licensee. This started a trend, and over the next several years more large companies wanted to install our software in-house. These companies included Abbott Laboratories, Nestle, Mars, Pet, Sun Oil, Bristol-Myers, and Clorox. We continued to do logistics optimization projects in the early 1980s for Basic American Foods and R&G Sloane Manufacturing, but the trend was clearly shifting toward in-house licensing of INSIGHT software. This trend reflected the existence of competent corporate planning staffs and management science professionals who wanted to acquire powerful

optimization software to use themselves to support their organizations' strategic and tactical planning activities.

In addition to SAILS consulting and licensing, INSIGHT had an increasing flow of custom optimization work through the 1980s. Indeed, by the mid-1980s roughly 70% of our revenues were from custom optimization work with 30% coming from licensing and consulting with SAILS. This custom modeling work came mainly from large companies with management science staffs who were trying to solve large and complex resource allocation problems. We often joked that INSIGHT was the stop of last resort when in-house modeling groups had tried everything else to solve their models and failed. These corporate management science professionals were familiar with the management science literature and found INSIGHT to be a company with outstanding representation in relevant research published in the top refereed journals. The results we were delivering to our clients and that were published in *Interfaces, Management Science*, and *The Journal of Business Logistics*, among other top-rated journals, were evidence that if you had been struggling with a tough resource allocation optimization problem without success, it was worth a call to INSIGHT.

Although we developed and implemented powerful optimization systems for capital budgeting and portfolio selection (GTE, Mobil Oil), petroleum dispatching (Chevron, Mobil Oil, Getty Oil), and airline crew scheduling (United Airlines), the greatest amount of our custom optimization work involved various aspects of production planning and scheduling (Basic American Foods, Clorox, M&M/Mars, Nabisco, Eli Lilly, Kellogg's, Iowa Beef Processors, Anheuser-Busch).

During the last half of the 1980s the licensing of SAILS for in-house company use increased steadily. The gap between license revenue and custom software development was narrowing. Then, in 1989, two large custom projects were launched which came to play a significant role in INSIGHT's already shifting emphasis toward licensing packaged products. These projects were the Global Supply Chain Model (GSCM) for Digital Equipment Corporation and the Heavy Products Computer Assisted Dispatch (HPCAD) system for Mobil Oil. Both of these projects were completed with great success and reported in *Interfaces*. A decision was made to put these two modeling systems into packaged form for stand-alone use on the increasingly powerful PCs of the mid-1990s.

In this same period, 1991–1992, we moved SAILS to the PC from the mainframe. PCs had finally become serious computing platforms where large optimization programs could be executed in reasonable times. Many clients continued to use SAILS on their mainframes, but there was rapidly increasing demand for "easier-to-use" and graphically appealing software which could be used independent of the corporate information system bureaucracy. The logistics analysts wanted to have their own models on their own machines on their desks. Our first PC-based SAILS client was GE Appliances in 1992. Many other PC implementations of SAILS quickly followed, and by the mid-1990s our revenue pattern had flip-flopped in that 70% was now from licenses for SAILS and our "new" packages, GSCM and SHIPCONS II, while 30% was from custom optimization work.

A significant phenomenon was becoming apparent by the mid-1990s which continued up until the time I retired as CEO of INSIGHT in 2003. As corporate America "re-engineered" to be more competitive in a global economy many large businesses, including our client base, reduced or eliminated their corporate management science and planning staffs. Those groups of management science and logistics planning professionals which had been our key contacts and users of our software in client organizations started disappearing. As a consequence more businesses started looking to buy turn-key solutions from outside sources. This was reflected in the move to buy enterprise resource planning (ERP) and supply chain management (SCM) suites. It was also reflected in more companies asking for consulting support to do logistics optimization projects, even when they had already licensed the SAILS software for in-house use. Ironically, this change in the environment for our services took us full circle back to where we began in 1978, using our proprietary software to conduct analyses for our clients. That pattern continued until my retirement in 2003.

6.8 Bumps in the Road

Every business has the same set of obstacles to overcome to achieve success: financial, technical, organizational, and market presence. We had all of those to be sure, but they were always met and overcome. However, we did have a consistent set of issues that caused frustration. The first source of frustration was the organizational inertia of some of our clients and potential clients. Notwithstanding the frequent admonishments in the management literature of the dysfunctional effects of organizational silos, we found that for most organizations those silos were well entrenched. The traditional functional divisions of manufacturing, finance, marketing and logistics viewed the world and their businesses from the narrow perspectives of their own divisions' best interests. This view of the world was reinforced, and indeed caused, by the compensation systems that existed in most client companies. As a consequence, when a cross functional analysis was done that looked at the business as a whole one or more of the functional divisions would view the results as detrimental to its own division's interests. The response, not surprisingly, would be to try to torpedo or discredit the analytical results and thus prevent the implementation of what we proposed as a way to increase the overall return on assets for the business.

A second source of frustration was the frequent inclination of client organizations to seek "simple solutions." We published an article on this very subject in *INTERFACES* in 1983, *The Myth of the Simple Model*. Although analyzing the entire logistics system of a major business organization was an inherently data-intense and complex undertaking, many potential clients wanted a methodology like spreadsheets that they understood. We often commented that an organization would rather accept an inferior or wrong solution than accept one they did not fully understand. Many potential client organizations did not have trained management

scientists who were comfortable with mathematical optimization technology. Logistics planners were uneasy trusting the results of a process they did not fully grasp. Consequently we sometimes lost assignments to competitors who offered simple heuristic or simulation approaches. The fact that we could demonstrate with examples that such "simple" approaches not only did not guarantee the best result but sometimes the wrong result, did not carry the day.

A third source of frustration was the emergence of Enterprise Resource Planning (ERP) systems. Those who adopted such systems became captives of what data were readily available in the ERP databases as well as warnings by the vendors of those systems that any analytical programs other than what they provided were incompatible with the ERP package. As the ERP systems were usually committed at very high levels in client organization, and for a different set of reasons than the support of logistics network analysis, we sometimes found ourselves excluded for fallacious reasons. We even had existing clients who had been using our software for years who had to spend large amounts of time and money to simply extract the data from ERP systems that had always been available in legacy systems that were displaced.

6.9 The View Ahead

As was noted earlier, the loss of management science professionals in many client and potential client organizations has continued. It seems the currency has been debased in far too many instances to the point that client organizations want grand consulting solutions using simple tools inadequate to the task of modeling highly complex logistics networks. Without the management science expertise in-house to adequately evaluate options offered, low-technology solutions are quite often preferred because they seem easy to understand.

On the other hand, for those organizations that do have the expertise to grasp the value of truly globally optimal solutions more complexity and richness are being incorporated into logistics network models, thus placing ever greater demands on the solution technology. Forward-looking companies want to consider "green supply chains" where energy consumption is a component of the optimal solution. Rather than the classic logistics network of producer–distribution center–customer, much more comprehensive logistics network models are sought: incorporation of raw material sources, marketing impacts of various configurations, seasonality of demand or materials, and multiple stages of production and conversion. Postponement strategies and inventory stratification and staging are increasingly looked at with comprehensive network design models. So we have two almost opposite effects occurring at the same time: organizations that will settle for "simple" methodology in a larger consulting context, and organizations with management science professionals who are demanding ever more capable solution technologies to handle far more complex model features than in the past.

6.10 In Sum

My 25 years applying management science not only to corporate America, as we started out to do, but to corporations and governments all over the world, has been stimulating, challenging, and I believe for our clients, quite profitable. We executed scores of client engagements and license support assignments for the top companies in the world. At one point in 2001 I counted 40% of the Fortune 50 as our clients and, excluding purely financial firms like banks and insurance companies, 45% of the top 50 companies in the Business Week Global 1000 were our clients. We had long observed that our clients were consistently among the most profitable firms in their industries. We believe we contributed to that, but more importantly it reflected that the top companies recognized the value of management science and modeling in making resource allocation decisions. Although many of our clients did not divulge to INSIGHT the magnitude of their ROA that resulted from using our software or from our consulting engagements, I am confident the savings in operating costs and asset reductions ran to the tens of billions of dollars. One client alone, Digital Equipment, reported savings in operating costs over 4 years of $1 billion and asset reductions of $400 million from decisions made based on the use of GSCM.

On the professional side, INSIGHT's staff has had a tremendous record of articles published in the top refereed journals in management science and logistics. Many of the scores of articles published by INSIGHT's staff members have described modeling work done with our clients. In addition to a prodigious volume of seminal articles in the professional literature there has been recognition of other sorts. INSIGHT clients were runners-up for the Edelman Prize on three occasions. INSIGHT staff members were frequent speakers at national conferences for management science and logistics as well as invited guest faculty for several university executive management programs. Several INSIGHT members have played prominent leadership roles in the top professional organizations for management science and logistics. Finally, and most significant, Geoffrion and Brown, two of the original members of INSIGHT, were elected to the National Academy of Engineering.

It was a wonderful 25 years spent with top-notch associates and loyal clients, many of whom became and remain good friends.

Chapter 7
Optimization Tradecraft: Hard-Won Insights from Real-World Decision Support*

Gerald G. Brown, Richard E. Rosenthal

"Thou shalt never get such a secret from me but by a parable."
Shakespeare, The Two Gentlemen of Verona

This paper honors the memory of deceased coauthor Richard E. Rosenthal

Abstract Practitioners of optimization-based decision support advise commerce and government on how to coordinate the activities of millions of people who employ assets worth trillions of dollars. The contributions of these practitioners substantially improve planning methods that benefit our security and welfare. The success of real-world optimization applications depends on a few trade secrets that are essential, but that rarely, if at all, appear in textbooks. This paper summarizes a set of these secrets and uses examples to discuss each.

Clients consult specialists because they have real-world problems to be solved. Clarifying a problem statement by talking with a client or, better, getting first-hand experience with the client organization is very different from reading a textbook case study. (However, some clients might feel that your success would threaten their jobs.) In this paper, we offer advice that we learned from completing hundreds of optimization-based decision-support engagements over several decades. These are hard-won lessons based on field experience. As a practitioner of our optimization art, you must obtain some experience beyond textbook coursework before these suggestions will make complete sense to you. Thus, you will not find this material highlighted in any textbook. Providing decision support in the real world is difficult because it requires that you deal with enterprise data systems, legacy procedures, and human beings who might not share your passion for making things better.

Gerald G. Brown
Department of Operations Research, Naval Postgraduate School, Monterey, California 93943, USA

* Reprinted by permission, Gerald G. Brown, Richard E. Rosenthal: Optimization Tradecraft: Hard-Won Insights from Real-World Decision Support, *Interfaces* 38(5), 356–366, 2008. Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 310, Hanover, MD 21076 USA.

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_7,

We receive many phone calls from colleagues and ex-students who are working with optimization. Sadly, too many of these callers do not extol the wonders of optimization; rather, they lament practitioner problems in getting things to work right. Unfortunately, this may have given us a distorted view of the issues we address here.

In this paper, we present our tradecraft in the topical categories that we have used to collect our lessons learned. Even if you are not a practicing optimizer, we suspect you will find insights here.

7.1 Design Before You Build

We have had an astonishing number of opportunities to address problems with optimization models that have been implemented, but are behaving badly (e.g., they are very hard to solve, too large to solve, or produce strange results) and are not documented. They have been built without a design!

Documentation must—not should, must—include these three critical components:

- A nonmathematical executive summary,
- A mathematical formulation, and
- A verbal description of the formulation (Figure 1).

A *nonmathematical executive summary* must answer the following five questions, preferably in this order (Brown 2004a):

- What is the problem?
- Why is this problem important?
- How would the problem be solved if you were not involved?
- What are you doing to solve this problem?
- How will we know when you have succeeded?

Express your executive summary in your executive sponsor's language, rather than in technical jargon. If you have trouble writing such a summary in less than five pages, you are not ready to proceed. The following tricks will make writing your summary easier and more effective:

- Have a nonanalyst read your executive summary to you, out loud,
- Ask this reader to explain your executive summary to you,
- Listen well, and
- Revise and repeat.

A *mathematical formulation* should include the following in this order (Brown and Dell 2007):

- Index use (define problem dimensions),
- Given data (and units),
- Decision variables (and units),

- Objectives and constraints, and
- (perhaps) a dual formulation.

Remember to define terms before using them. The earliest definition of such a standard formulation format appears in Beale et al. (1974). To distinguish inputs from outputs, adopt a convention such as using lowercase for indexes and data, and uppercase for decision variables.

A *verbal description of the formulation* (Figure 7.1) explains, in plain English and in your executive sponsor's language, what each decision variable, objective, and constraint adds to the mathematical model. It gives you the opportunity to define what the mathematics means and why each feature appears in your model. Avoid literally translating mathematics into English. For example, avoid saying "the sum of X over item subscripts i must be no more than m for each time subscript t." Instead, say "the total production of all items must not consume more raw material than will be available in any year." Do state and justify any simplifying assumptions (some examples include "our planning time fidelity is monthly, with a 10-year planning horizon," and "we allow fractional production quantities of these large volumes").

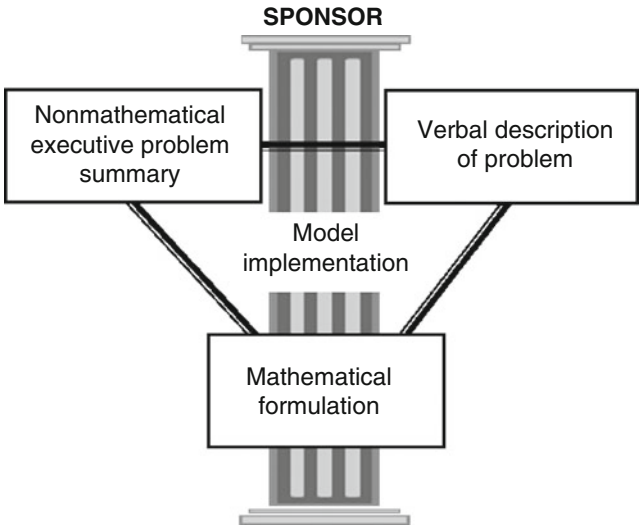


Fig. 7.1 The model sponsor will only likely see the nonmathematical executive problem summary and the verbal problem description. The actual model implementation must be embedded with these two essential documents and with a mathematical formulation. In our experience, there is no substitute for any of these components.

7.2 Bound All Decisions

Bounds restrict the domain of every decision. An unbounded variable does not exist in our real, OR analyst's world. Establishing bounds for each decision variable is a trivial concept that is often ignored. While any reasonable optimization solver will

do this automatically, the solver cannot tell you that its analysis is based on bogus data or missing features in your model. If you manually apply simple ratio tests (e.g., "If I had all the steel the world produced this year, how many automobiles could I build?") and get ridiculous answers (e.g., "2.1 autos," or "10 trillion autos"), you have discovered an error either in the data or in the description of the manner in which automobile production consumes steel. These conversions reveal an erroneous steel consumption rate per auto or a constraint that has no influence on your model; thus, you can jettison them.

Do you remember all the formal "neighborhood" assumptions that underlie your optimization method? Taylor's theorem makes any continuous function appear linear if you bound your decision neighborhood tightly enough. All your costs and technology likely exhibit nonlinear effects across widely varying magnitudes; however, they might not exhibit the same effects over a small neighborhood—the domain for which you are planning.

It is easier to branch-and-bound enumerate models with integer variables if the bounds on the integer variables are as tight as possible. This is worth addressing before you try to solve large models. If the tightest bounds that you can state permit a "large" integer domain, relax the integrality requirement and round the continuous result to the nearest integer. The inaccuracy that rounding inflicts will be no worse than one divided by the final value of the variable.

Bounding all your decision variables pays an unexpected bonus. Pull out your favorite optimization textbook and look at the basic theorems that might have seemed so hard in class. Notice how much mathematical lawyering becomes superfluous when you rule out the unbounded case. Voila!

7.3 Expect Any Constraint to Become an Objective, and Vice Versa

Important planning models almost always exhibit multiple, conflicting objectives. Get familiar with a "weighted average objective," and what it really means. Learn about "hierarchical (i.e., lexicographic) objectives," and how to coerce off-the-shelf optimization software into following your hierarchy. For example, you might maximize the highest-priority objective, and then add a constraint on this objective to maintain this performance in all subsequent solutions. Repeat this process with each lower-priority objective until these successive restrictions have addressed your entire hierarchy, or your model is so overconstrained that further restriction would be pointless. Using some algebraic modeling languages, you can automate all of this as a single model excursion.

You can see that there is a continuum (sic) between weighted objectives in a single monolithic model, and strictly hierarchical ones in a sequence of successive restrictions. It is possible to force hierarchical results by using wide-ranging values for weights; however, you might regret the attempt. Take care to use your model-generation logic to control a hierarchal-solution sequence, rather than try to

force your optimization model to make this asymptotic transition from finite weights to the infinite weights required to render absolutely lexicographic results. Floating-point numerical errors increase in direct proportion to the relative magnitude of the terms in your additive weighted objective. You might be able to express such an objective; however, your solver will not see what you intend to be lower in the objective hierarchy.

In one of our engagements, we dealt with an extreme case with 14 objectives, each weighted at least an order of magnitude more than its predecessor in the weighted hierarchy. This was not a pretty numerical experience for the solver.

Expert guidance from senior executives frequently filters down to modelers as constraints (i.e., orders). In our experience, constraints deriving from literal interpretation of such guidance inevitably lead to an infeasible planning model. Discovering what can be done changes your concept of what should be done. This leads you to "aspiration constraints," a situation in which you determine how much of something you can maximize in isolation; you can then write a constraint saying, for example, "I'll settle for 90 percent of this isolated maximum." If you work with your senior sponsors using these simple methods, you will be able to guide them to give you better advice. As OR analysts, we may think that our job is to give advice; however, our real objective is to help our sponsors to make the right decisions.

Much of the relevant literature advises us on how to deal with multiple objectives. It does a nice job of defining and explaining concepts, such as pareto-optimality. However, simple ideas usually work best.

7.4 Classical Sensitivity Analysis Is Bunk—Parametric Analysis Is Not

Blind application of dual values, right-side ranging, and other textbook tricks offer little useful advice on how the solution will respond as the inputs *all* change. Even for the few models that are continuously linear, classical textbook sensitivity analysis is rarely useful. Some of the best off-the-shelf mathematical modeling languages and solvers do not support such analysis. We professors love to teach this "stuff." We will continue to teach it because it conveys lessons on the foundations of our optimization methods, and on how to interpret the quantitative (how much to do) and qualitative (what to do) influence of restrictions and relaxations.

However, in the real world, *plan on solving many model excursions*; do not hesitate to try this approach because "it may take weeks to complete." In the past 15 years, improvements in linear program (LP) solvers and, in particular, in integer linear program (ILP, aka MIP) solvers and their controls have improved performance by a factor of at least 10,000 *independent of the much faster speeds of newer computers*. Some in our profession, especially the senior, experienced professors and textbook authors, still recall overnight batch processing of mathematical programming system (MPS) tapes. This is not a fond memory; therefore, our advice is simple—"get over it." All you need today is a reasonably endowed desktop or

laptop computer. For almost any modeling engagement, we can expect to set up an optimization model that allows us to express a question and get an answer while our sponsor still remembers the question.

7.5 Model and Plan Robustly

Ensure that your model considers alternative future scenarios and renders a robust solution. There are many ways to capture this in your model; all boil down to arriving at a single plan that, if applied to any of your scenarios, solves that scenario with acceptable quality, and which you can express as some combination of feasibility and optimality.

In the military, we plan for what is possible, not what is likely; therefore, we seldom employ random variables to represent the likelihood of each alternative future. We use simulation to make quantitative (perhaps random) changes to data elements; however, we rarely randomly sample qualitative future changes. Senior planners use judgment to arrive at what they think is a fully representative set of deterministic scenarios. While there could be many theater-war plans, we normally only have one chance per year to request what we need to prepare for all of them.

We pay attention to the current defense-planning guidance. As we develop our model, we try to address the sponsor requirements. For example, suppose that our guidance is to fight and win one engagement while suppressing another, and then to fight the other and win it. If we do not have the option of selecting our favorites of 20 available war plans for such potential engagements, we might have to plan for 20-times-19 permutations of engagement pairs.

You might not be able to develop a plan that addresses all scenarios; thus, you could be motivated to search for a worst-case plan, which will distort your results. It is better to convey truthful insights to your sponsor than to delude yourself with baseless optimism. From the full scenario set, we can devolve to, for example, meeting a maximum subset of scenario requirements, or maximizing some gauge of scenario fulfillment. Whatever plan you select, do your best to document with exquisite clarity your assumptions and compromises that differ from the overarching defense-planning guidance. Despite apocryphal tales of the demise of analysts bearing bad news, an OR analyst who uses diplomatic, unambiguous language and careful analysis to deliver bad news will be a hero.

We seek the worst case among a reasonable set of outcomes that we control because that is what we are obligated to worry about and defend against. There are many commercial analogs to this advice. We find little to distinguish private-sector competition from military planning.

7.6 Model Persistence

Optimization has a well-earned reputation for amplifying small changes in inputs into breathtaking changes in advice.

Decision-support engagements typically require many model excursions, followed by analysis, followed by revisions and more model runs. When we have invested heavily in analyzing a legacy scenario, and must make some trivial adjustment to attend to some minor planning flaw, the last thing we want is a revision that advises major changes. This is always an issue with rolling-horizon models; it also arises when you make iterative refinements to a static model.

If your model is unaware of its own prior advice, it is ignorant. You can expect annoying turbulence and disruption when solving any revision of a legacy model. Any prescriptive model that suggests a plan, and, if used again, is ignorant of its own prior advice, is free to advise something completely, needlessly different. This will surely cost you the faith of your sponsor. Sometimes, there are many nearly optimal policies; however, if you have already promulgated one of these, it is now a legacy-planning standard that is worth trying to preserve.

Persistence means "continuing steadily in some course of action." This is exactly what we do with long-term optimization-based decision-support engagements. We must successively meld our sponsor's expert judgment with our model's optimal advice.

It is easy to add model features that limit needless revisions. To do this, you need to *make a published legacy solution a required input*, and then add model features to retain attractive features or limit needless revisions of this legacy. These persistent features might include the following (Brown et al. 1997):

- Do not change this legacy resource consumption by more than 2 percent,
- Between this legacy solution and any revision, add (or delete or change) no more than three of the binary options in this set,
- Do not change X unless you also change Y.

We give our students a handout showing them how to state integer linear constraints that express the ubiquitous logical relationships required in decision support (for example, for binary options A and B, "A only if B," "A and B, or neither," "A or B, but not both," or "A or B, or both"). We also show them how to state persistent guidance for revisions (because this information rarely appears in textbooks). For example, the Hamming distance between a legacy vector of binary decisions and a revision counts only the bit-wise number of changes. To solve a sequence of revisions, you can use constraints either to limit the number of revisions; in cases in which you are looking for a set of alternative courses of action to present to your sponsor for subjective evaluation, you can force diversity of each revision from *any* legacy solution (Brown and Dell 2007).

The literature suggests widely that in a facility location, for example, one should use a binary variable to represent each close-open decision with a fixed cost inflicted when we choose open. *We rarely get to apply this in the real world* because each facility might be in one of several states (e.g., open, open but idle, mothballed, closed, or disposed); the real problem is to decide which state transitions are best for the client. In even the simplest case, we have preexisting legacy facilities and their states and we choose revisions of those states; in these revisions, each before-after state pair has its own distinct, fixed transition cost. Multiperiod planning requires a

binary variable for each state transition and a constraint to force choice of only one transition per decision.

Solution cascades (Brown et al. 1987, p. 341) solve a window of active constraints and variables moved over, e.g., time, fixing the values of each variable as the value determined when it was last in a window, for several reasons. For example, omniscient long-term optimization models sometimes are too clever about anticipating the distant future; we prefer more realistic time-myopic planning. We can also use persistent cascades to incrementally revise a plan locally while preserving its overall scheme. Sometimes, the cascade subproblems are much easier and faster to solve in large numbers than the seminal, monolithic model.

We also wonder why our literature pays scant attention to end effects. When we plan on using periodic state reviews over a finite number of planning periods, how do we plan to leave our system at the end of this planning horizon? There may be industry rules of thumb or policies on the admissible state of your enterprise (e.g., always have sufficient supply on hand to satisfy the next 90 days of demand). Lacking such guidance, we often plan further into the future than the planning horizon requires because we want to get some realistic representation of the actions up to exactly the end of the planning horizon (and discarding the further future results) (Brown et al. 2004).

7.7 Pay Attention to Your Dual

A conventional linear program equality constraint has an unrestricted dual variable that we can interpret as "this is how much it would be worth to relax this constraint by one unit."

An *elastic* linear program equality constraint uses a linear penalty per unit of violation below (or above) its minimal (or maximal) range. Allowing this constraint to be violated below (or above) either range at some finite penalty cost-per-unit violation bounds its dual variable (i.e., "this is the most it is worth to me to satisfy this constraint; otherwise I'll violate it, pay this penalty, and deal with the consequences"). There is no such thing as an infinitely valuable constraint. Decision makers get paid to deal with infeasibilities and cannot rule them out in the real world.

When you convince your sponsor to work with you to state each constraint with a well-planned penalty for its violation, you have enormously enhanced your control and understanding of your decision-support model. Remember that a phone call beats a clever planning method every time. That phone call could be between you and your sponsor, or between the sponsor and a supplier, superior, or even the IRS. A written problem description or model statement could never have the level of impact that relaxing exasperating restrictions does. Managers are paid to make these calls and deal with infeasibilities.

Elastic constraints provide another surprise bonus: integer linear programming is much easier to deal with when you know a priori that every candidate integer

solution in an enumeration is, by definition, admissible (i.e., satisfies the constraints, albeit perhaps with some penalties). In addition, if you set your elastic penalties carefully, you will be rewarded with remarkable improvements in linear-integer solution quality and solver responsiveness.

If you have a linear program, or can relax to one, state its dual. If you cannot write an abstract of the meaning of this dual, if you cannot interpret your dual at all, or if your dual is nonsense (e.g., unbounded or infeasible), your primal problem is ridiculous. OK, this is strong language. Amend this to read "your primal problem needs more attention before you are ready to use it."

Consider this example of a simple maximum-flow model that we have used for military planning and, since 9/11, for planning homeland defense. It includes a source node, a destination node, and a capacitated, directed network through which we wish to push the maximum-flow volume from source to destination. Write this primal linear program and solve it. Now, recover the dual solution. Admire these dual values and note that each arc on a minimum cut is distinguished by two incident dual values that differ. *If you want to attack this maximum-flow network and can cut these arcs, you have decapitated it*.

Interpreting linear programming duals is the foundation of decomposition (Brown et al. 1987) and the bilevel defender-attacker or attacker-defender models (Brown et al. 2006).

7.8 Spreadsheets (and Algebraic Modeling Languages) Are Easy, Addictive, and Limiting

OK, we have a new problem; we need a quick answer; we need database support for model development and cataloging solutions; and we need a graphical user interface that supports ad-hoc analysis and graphical output. Thus, we must either spend a long time and a small fortune developing a purpose-built graphical user interface or use our off-the-shelf office software suite.

Spreadsheets with embedded optimization solvers are inviting. Even executive sponsors likely know how to bring up a spreadsheet; therefore, you will gain immediate acceptance by adopting this familiar "look and feel" standard. In addition, you will be able to catalog and display a spreadsheet solution immediately by using the tools you use in your integrated office software suite daily.

However, spreadsheets support only two-dimensional views (and pivot tables) of many-dimensional models; they exhibit "dimensional arthritis"—they can support a many-dimensional model; however, they do not do it easily or naturally (Geoffrion 1997).

We get many calls from spreadsheet users who wonder why their optimization results either take forever or generate incorrect results. One of the first questions that we ask is, "how much did you pay for the solver you used?" Consider spending a few thousand dollars per seat on a well-known, off-the-shelf, supported, and documented, commercial-quality optimization package. In addition, before you commit

to using any solver, check the credentials of the optimization software provider and verify how you will get help if you have problems.

Modeling languages are crafted to accommodate multidimensional models; they feature interface links to all contemporary database, spreadsheet, and presentation managers, and make great prototypes. However, even if a prototype works and gains acceptance, the modeling language used for prototypic implementation might not make a good decision-support tool. Some modeling languages isolate models from off-the-shelf commercial solvers. They do not provide good support for large-scale, indirect-solution methods (for example, column generation or decomposition). If you are working on an important problem, why would you jettison 40 years of experience in solving it well, and, instead, simplify and aggregate away essential details merely to be able to mechanically generate and solve problem instances?

The transition from hasty prototype to production-model generator and interface is not easy. However, in our experience, the results always justify the investment. The use of a commercial-quality optimization package could reduce your model-generation and solution times from hours to just seconds (Brown and Washburn 2007).

7.9 Heuristics Can Be Hazardous

A heuristic—whether a simple rule of thumb or a well-known local search method—is so easy to explain and implement that we are often tempted to use one in lieu of more formal methods. Heuristics might not require optimization software and might offer a tantalizing first choice to quickly assess a "common sense" solution. However, heuristics should rarely be your first (or only) choice. Geoffrion and Van Roy (1979) offer some simple, exquisite examples that they have used with executives to show how blind adoption of common-sense heuristics can bring you grief.

We can also develop bounds on the best solution possible, although this is not as much fun to do as building a solution-seeking method. Without some similar bound, our advice is of unknown quality. This quality certification is important: *a bound on the value of the best possible solution is just as important as the best solution you have*.

A mathematical optimization model takes longer than a heuristic to develop, and perhaps to solve; however, it can provide a bound. We develop models of relaxations of very hard problems merely to recover the bounds that they provide. Lacking a trustworthy assessment of the quality of your advice, you are betting your reputation that nobody else is more scrupulous or just plain luckier than you are.

While publishing a bound with your solution is the right thing to do, there is a risk. We have been told: "Hey, you're leaving money on the table!" Well, maybe we are and maybe we are not. At least, we are honest about the possibility.

The *interval of uncertainty* is what we call the interval between the value of a solution and a bound on the value of the best-possible solution (various sources exist, including integrality gap, decomposition gap, Lagrangean gap, and duality gap).

When you compare two alternative scenarios, you can be absolutely sure about the winner if the two intervals of uncertainty are disjoint, no matter how large each of these intervals is. Realizing this, you can work only hard enough to find a distinguishing difference—and no harder.

We have also been in a private-sector competition in which our heuristic competitors wrote the sponsor and said, "these guys admit their solutions may not be right." Boy, they thought they got us there, didn't they? To this, they responded "but, our method gets better solutions the longer you run it." This reminds us of the difference between "known unknowns" and "unknown unknowns." We can work with the former; we get nightmares from the latter. While a heuristic might suggest a provably better plan than the plan the enterprise is using currently, you will never know how much more you might have discovered. Would we implement a solution with no quality assessment? No, thanks.

We have also been told (sigh, and have read in the literature) that "this ILP is NP-hard, so we use a heuristic." Please. Even if (ahem) you prove that your ILP is NP-hard (an essential reduction proof that is still absent from our literature too frequently), this only means it is as hard as many other problems that are routinely and reliably solved to good tolerance. How much better is a heuristic with polynomial run time than a bounded ILP enumeration, which benefits from hundreds of years of research and experience by our optimizers? In addition, is the heuristic really any faster?

The simplex method has been criticized for its exponential worst-case run time on polynomially complex linear programs. Given its excellent average performance on an immense diversity of real-world linear programs, the worst-case run time limit is a poor excuse to adopt an alternative solution method. We have a good idea of the classes of problems for which the simplex method works well.

We prefer to solve any model that we can, even approximately, using conventional mathematical optimization and the best software we have. If we convince our client that our suggested planning tool is worthy, software that costs a few thousand dollars per seat should not be a problem.

In cases in which the cost per seat would be too high to distribute the best software we have, or the number of seats required is necessarily high, and the model admits a heuristic solution, we try to develop a heuristic. Using our best software, we test empirically to assess performance. If we distribute the heuristic, we maintain a backup with our more-expensive software to objectively assess any curious performance in the field. At the Naval Postgraduate School, this means that we must maintain computers and software at various classification levels in appropriately secured facilities. While this requires a significant investment in hardware and software, it is essential to providing a safety net for fielded heuristic solvers.

We have encountered other obstacles both in the government and in the private sector with "enterprise standard" computers that are not allowed to run "foreign" executables and "exotic" applications, such as our optimization models. For example, Navy Marine Corps Intranet (NMCI), which governs 351,000 computers, is the largest standardized internal computer network worldwide (Electronic Data Systems 2006). Presumably this standardization has had benefits for "one size

fits all" IT support. However, it has been a continuing headache to us. We cannot afford to have each of our models "vetted" and "approved" (a process that takes many months and many thousands of dollars) for NMCI. Accordingly, we have developed heuristics that can run, for example, with Visual Basic within Microsoft Excel on a standard NMCI computer. We have also developed applications that run exclusively on a universal serial bus (USB) drive that can be connected to a NMCI computer.

We have also had to purchase computers, install our applications, and ship these to our clients. We refuse to confirm or deny where these clients serve, or if they also have their own private computers to do mission-essential work outside of NMCI. We do whatever is necessary to complete our missions.

Perversely, one of the most influential arguments for heuristics, and against excellent, off-the-shelf commercial optimization solvers, is the Draconian license managers of these solvers, which treat paying clients like criminals. We have seen many cases, in academe and in industry, where a good solver would have helped; however, it was rejected because of the sheer IT burden it would cause—that of struggling with optimization-provider sales persons, computer-specific, immobile license keys, and license-manager hassles.

7.10 Modeling Components

Models usually exhibit a variety of functional components that express different aspects of the modeled enterprise. Observe how this enterprise is organized and mimic this with your model. For example, when production plans influence financial plans, link these components with "passenger variables" (a passenger variable does not change the degrees of freedom in your model because it is defined by an equation) that isolate and highlight this communication between components. Choosing passenger variables deserves some care; you are trying to capture how the connected enterprise components communicate with each other.

You might think that cluttering your model with superfluous passenger variables and defining equations makes the resulting, larger model harder to solve. Fortunately, solvers employ "presolve" features that quickly identify "rank-one" algebraic redundancies (e.g., those that are identifiable without substituting more than one variable for its defining equation); remove them from the model before you solve it; then substitute them back in when you have completed the solution.

Incremental development of components offers an added benefit. During this phase of development, you need only work with representatives of the enterprise component that you are currently modeling; thus, you can focus without distraction on the lexicon, operation, fidelity, and key issues to capture. Better yet, you can arrange each component to be optimized in isolation during development and testing. Fix or constrain the passenger variables linking to other components, run the component alone, and unwind any mischief that appears in this localized exercise.

7.11 Designing Model Reports

Design model reports to match those that planners are already using.

It is not unusual to spend as much time in reporting as in modeling. For example, if you find that a Gantt chart is a key display that manual planners use, mimic it. If your model has significance for the enterprise, i.e., if your optimized plans can materially change profitability, plan on producing a set of operating statements. Such statements might contain a cash flow report, income statement, and balance sheet, including the most important gauge—return on owners' equity. This is difficult work because preparing such statements requires much enterprise operating data that you would not otherwise need. The payback for doing this foundation work is two-fold: you gain a deeper appreciation for where and how your model can influence the enterprise, and these synthetic reports will get the attention of your sponsor.

For example, if your advice might require raising significant amounts of funding (e.g., by borrowing, selling stock, issuing bonds, or diverting funds from other uses), the sources, methods, and forecast consequences of such fundraising are essential features of your model. If your objective is earnings per share, and both earnings and number of shares are discretionary, you have a ratio of decision variables that you might (or might not) be able to back out algebraically into a linear (sic) integer program. While this greatly complicates your modeling, it is essential to your reporting.

To our knowledge, the earliest example of such operating-statement reporting appears in paired papers by Bradley (1986) and Geoffrion (1986), who advised the board of directors of General Telephone and Electric (GTE) Corporation how to commit huge capital improvements with substantial impact on corporate results. Contributions by their GTE cohorts in this modeling project accompanied these papers. These authors generously provided us with all their historical client notes and model source code; we have dissected these and reapplied their methods.

We have had the distinct pleasure of working with both closely held companies and sole proprietorships. These owners quickly grasped optimization and its nuances, including integrality gaps, duality gaps, model fidelity, and uncertainty. Because their own money is at stake, they really engaged with the details and valued these operating statements. We have also had experience with scrupulously run, publicly held corporations; they also valued operating-statement outputs, but with not with the level of intensity of private entrepreneurs.

An added advantage accrues from reporting in terms of operating statements. The managers of various "stovepipes" (i.e., enterprise components that are strongly intraconnected, but weakly interconnected) in the enterprise can see their business component and its interaction with others. This provides a level playing field among these managers, and encourages them to plan, negotiate, and speak in a common language. We have seen cases where, for example, marketing wants to make its quarterly "numbers" for incentive bonuses, finance seeks goals that are stated in terms of float, accounts receivable currency, and cash-versus-debt positions, and manufacturing strives to meet production-standard goals. This is akin to the fable of blind men each touching one part of an elephant's anatomy, and guessing what the

animal looks like. If you gather these managers in the same room and ask them to look at the same integrated operating reports, wondrous insights will follow.

Optimization also enables the generation of reports that management might not have known were possible. For example, it is easy to embellish a customary demand-fill rate report with an estimate of the total landed profit (or loss) accruing from those sales. Wow, this gets attention!

Design model outputs that are directly useable as model inputs. In practice, we frequently repeat model applications to iteratively revise our advice with small changes.

7.12 Conclusion

You may ask "why aren't these simple topics part of basic optimization course work?" We have been asked this before, and respond: "where were you when these pages were blank?" These ideas may be simple; however, we know of no other source of instructional materials that addresses these real-world concerns.

While many analysts have successfully applied optimization to real-world problems, few will admit the failures and false starts that too frequently delay a planning project. For example, INFORMS Edelman presentations include some very impressive results; understandably, however, they rarely discuss the failures that occur on the path to completion. You might seek out these authors to learn, as we have, that the topics we report here are ubiquitous.

We have invested heavily to incorporate these principles into our graduate courses. In our program, each student is part of a group; the students attend a tightly coordinated, lengthy sequence of optimization core classes as a cohort. Thus, we have the luxury of getting to know and teach them individually and as a group over an extended period. While we have had some success in helping them to understand the material, it is not at a sufficiently high level. We have concluded that the only way students will appreciate the value of some of our advice, which might admittedly be tedious to implement, is through experience.

Accordingly, we try to convey these ideas to our military-officer students using both humorous, self-deprecating case studies of our past peccadilloes and homework exercises. However, we also realize that this will not make much of an impression until the student has had some seasoning. We include a continuing, evolving copy of this document in our course materials; we also give each graduate a "lifetime money-back guarantee" to call us later, admonishing them to have this document in hand when they do (Rosenthal 2007).

Suffice to say we have seen the same problems arise scores of times, even for very experienced operations researchers; we have cataloged some in this paper, along with our prescriptive cures.

We wish you the best of luck in helping us to extend our reach with prescriptive optimization-based decision support to make our world better and more secure.

Acknowledgments

This paper derives from decades of modeling engagements, many of which were exigent exercises assisting colleagues and past students. Successful rescue drills earned us a reputation, which led to an invited plenary tutorial at a Military Operations Research Society meeting (Brown 2002) and another at an INFORMS practice meeting (Brown and Rosenthal 2005). Along the way, we were asked to publish a "how to" guide for documenting optimization (Brown 2004a). Kirk Yost encouraged an intermediate "secrets to success" publication (Brown 2004b). In this paper, we focus on improving models that are *correctly* formulated. By collecting a rogues' gallery of examples that frequently lead to confusion (Brown and Dell 2007), Rob Dell helped us hone these topics and isolate the most common mistakes leading to *incorrect* formulations. We are aware that the references we cite here are insular. This is not an oversight. Our advice is so opinionated, we hesitate to implicate others. We also want to present a self-consistent, unified view of our complicated topic. These references are postcards home from a life journey in optimization. We credit our close colleague, Art Geoffrion, for his many insightful observations about the conduct of decision-support engagements (Geoffrion 1976a, b; Geoffrion and Van Roy 1979; Geoffrion and Powers 1980; and Geoffrion 1986, 1997). Most of all, we are grateful to so many students who have confronted real-world problems using the optimization tools we teach, and have claimed the "lifetime money-back guarantee" that we grant each of them to come back at us and complain that "neither my textbooks nor my notes from our courses explain this." *You students were right*. We fixed this with each of you and learned a lot along the way. We thank each of you. (And, every one of Distinguished Professor Rosenthal's many such warranties, public and personal, will be honored by me, and by my colleagues. Just get in touch with us.)

References

1. Beale EML, Breare GC, Tatham PB (1974) The DOAE reinforcement and redeployment study: A case study in mathematical programming. In: Hammer PI, Zoutendijk G (eds) Mathematical programming in theory and practice. Elsevier, New York, 417–442
2. Bradley GH (1986) Optimization of capital portfolios. Proc. National Comm. Forum 86:11–17
3. Brown GG (2002) Top ten secrets for successful application of optimization. Military Oper. Res. Soc. Annual Meeting, Ft. Leavenworth, KS, June 19
4. Brown GG (2004a) How to write about operations research. PHALANX 37(3):7ff
5. Brown GG (2004b) Top ten secrets to success with optimization. PHALANX 37(4):12ff
6. Brown GG, R. F. Dell. 2007. Formulating linear and integer linear programs: A rogues' gallery. INFORMS Trans. Ed. 7(2, January)
7. Brown GG, Rosenthal RE (2005) Secrets of success with optimization. INFORMS Practice Meeting, Palm Springs, CA, April 18
8. Brown GG, Washburn AR (2007) The fast theater model (FATHM). Military Oper. Res. 12(4):33–45
9. Brown GG, Dell RF, Newman AM (2004) Optimizing military capital planning. Interfaces 34:415–425

10. Brown GG, Dell RF, Wood RK (1997) Optimization and persistence. Interfaces 27:15–37
12. Brown GG, Graves GW, Honczarenko MD (1987) Design and operation of a multicommodity production distribution system using primal goal decomposition. Management Sci. 33:1469–1480
12. Brown GG, Graves GW, Ronen D (1987) Scheduling ocean transportation of crude oil. Management Sci. 33:335–346
13. Brown GG, Carlyle M, Salmerón J, Wood K (2006) Defending critical infrastructure. Interfaces 36:530–544
14. Electronic Data Systems (2006) EDS signs NMCI contract extension to 2010. Retrieved April 24, 2008, http://www.eds.com/news/news.aspx?news_id=2905
15. Geoffrion AM (1976a) The purpose of mathematical programming is insight, not numbers. Interfaces 7(1, November):81–92
16. Geoffrion AM (1976b) Better distribution planning with computer models. Harvard Bus. Rev. 54(4, July–August):92–99
17. Geoffrion AM (1986) Capital portfolio optimization: A managerial overview. Proc. National Comm. Forum 40(1):6–10
18. Geoffrion AM (1997) Maxims for modelers. Retrieved April 18, 2008, http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm
19. Geoffrion AM, Powers RF (1980) Facility location analysis is just the beginning. Interfaces 10(2, April):22–30
20. Geoffrion AM, Van Roy TJ (1979) Caution: Common sense planning methods can be hazardous to your corporate health. Sloan Management Rev. 20(4, summer):31–42
21. Rosenthal RE (2007) It's more than a job or an adventure. OR/MS Today (August):22–28

Part II
A Long View of the Future

Chapter 8
Challenges in Adding a Stochastic Programming/Scenario Planning Capability to a General Purpose Optimization Modeling System

Mustafa Atlihan, Kevin Cunningham, Gautier Laude, and Linus Schrage

Abstract We describe the stochastic programming capabilities that have recently been added to LINDO application programming interface optimization library, as well as how these stochastic programming capabilities are presented to users in the modeling systems: What's*Best*! and LINGO. Stochastic programming, which might also be suggestively called Scenario Planning, is an approach for solving problems of multi-stage decision making under uncertainty. In simplest form stochastic programming problems are of the form: we make a decision, then "nature" makes a random decision, then we make a decision, etc. A notable feature of the implementation is the generality. A model may have integer variables in any stage; constraints may be linear or nonlinear. Achieving these goals is a challenge because adding the probabilistic feature makes already complex deterministic optimization problems even more complex, and stochastic programming problems can be difficult to solve, with a computational effort that may increase exponentially with the number of stages in the "we, nature" sequence of events. An interesting design decision for our particular case is where a particular computational capability should reside, in the front end that is seen by the user or in the computational engine that does the "heavy computational lifting."

8.1 Introduction

We describe the stochastic programming (SP) capabilities that have recently been added to LINDO API (Application Programming Interface) optimization library, as well as how these SP capabilities are presented to users in the modeling systems: What's*Best*! and LINGO. SP, which might also be suggestively called Scenario Planning, is an approach for solving problems of multi-stage decision making under uncertainty. In simplest form SP problems are of the form: we make a decision,

Mustafa Atlihan, Kevin Cunningham, Gautier Laude
LINDO Systems Inc., Chicago, IL, USA

Linus Schrage
University of Chicago, Chicago, IL, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_8,

then "nature" makes a random decision, then we make a decision, etc. An underlying theme of our design of SP capabilities is, what are the features that are needed to make SP both (a) easy to use for relatively unsophisticated decision makers, but nevertheless (b) a powerful and useful tool. A notable feature of the implementation is the generality. A model may have integer variables in any stage. Constraints may be linear or nonlinear. Achieving these goals is a challenge because (a) adding the probabilistic feature makes already complex deterministic optimization problems even more complex and (b) SP problems can be difficult to solve, with a computational effort that may increase exponentially with the number of stages in the "we, nature" sequence of events. An interesting design decision for our particular case is where a particular computational capability should reside, in the front end that is seen by the user or in the computational engine that does the "heavy computational lifting."

8.1.1 Tribute

When we were designing the LINGO and What's*Best*! modeling systems in the 1980s, we benefited substantially from interactions with and from reading the papers of Art Geoffrion. In particular, as Art was writing the paper on indexing in modeling languages (Geoffrion [14]), we had regular interactions with him. The set handling capabilities of LINGO were much improved as a result. Art provided a general philosophy of modeling as outlined in his papers such as his "Insight, not Numbers" paper (Geoffrion [12]), and his "structured modeling" papers, see Geoffrion [13]. We found these papers very useful in providing general direction in designing a modeling system. Internally, we succinctly referred to these papers and their author as "The Art of Modeling." At one point in our discussions, Art made the comment that "One man's parameter is another man's variable." This particular comment affected the design of LINGO in two ways. In the declarations section of LINGO, a numeric attribute of a set element does not receive a type declaration such as *parameter*, *variable*, or *integer*. So, (a) an attribute becomes a variable only as a result of not being set to a value as part of data input and (b) a variable is declared integer or not as part of the model statements. Thus, in a multi-stage planning model we may want Produce(1) and Produce(2) to be restricted to integer values, whereas it may be convenient to allow the later Produce(3), Produce(4), etc. to be allowed continuous. In our design of SP capabilities for a modeling system, we have tried to remain true to what Art taught us.

8.2 Statement of the SP Problem

SP is concerned with solving multi-stage problems of decision making under uncertainty. An important concept in these problems is that of a "stage." Various

researchers have used various definitions of a stage. We have found the following description of SP and the role of a stage useful:

(0) In stage 0 we make a decision, e.g., how much to order, taking into account that later,
(1) At the beginning of stage 1, "Nature" makes a random decision, e.g., demand,

1.a) At the end of stage 1, having seen Nature's decision, as well as our previous decision, we make a decision, e.g., order some more, taking into account that ...

(2) Later, at the beginning in stage 2, "Nature" makes a random decision, etc.
...
n) At the beginning of stage *n*, "Nature" makes a random decision, and

n.a) At the end of stage *n*, having seen all of Nature's *n* previous decisions, as well as all our previous decisions, we make a decision.

Thus, a stage is defined as an ordered pair (random event, decision). Stage 0 is special in that there is no random event. The last stage may be special in that there may be no terminating decision. In some settings, e.g., Markov decision processes, one may be interested in problems with an infinite number of stages. We are here interested only in problems with a finite number of stages. We also assume that we are dealing with an indifferent nature, i.e., Nature's random decisions do not depend on our decisions, although Nature's decision in stage *n*, may depend on Nature's decisions in earlier stages. If there are only a finite number of outcomes (which is true computationally) for nature at each stage, then it may be helpful to visualize the process by a tree, as in Figure 8.1.

8.2.1 Applications

SP has been applied, or proposed, for a wide range of problems. A collection of examples appear in the book edited by Wallace and Ziemba [25]. Specific examples therein are fleet management, production planning, metal blending, mortgage refinancing, electricity generator unit commitment in the face of uncertain demand, and telecommunications planning over unreliable networks. Additional examples elsewhere are financial portfolio planning over multiple periods for insurance and other financial companies, in the face of uncertain prices, interest rates, exchange rates, and bankruptcies, see Carino and Ziemba [5]; capacity and production planning in the face of uncertain future demands and prices, Eppen, Martin, and Schrage [8]; fuel purchasing when facing uncertain future fuel demand and prices, Knowles and Wirick [18]; metal blending in the face of uncertain input scrap qualities, Gaustad et al. [11]; fleet assignment: vehicle type to route assignment in the face of uncertain route demand, Dantzig [6]; and hydroelectricity generation in the face of uncertain rainfall, Pereira and Pinto [22];

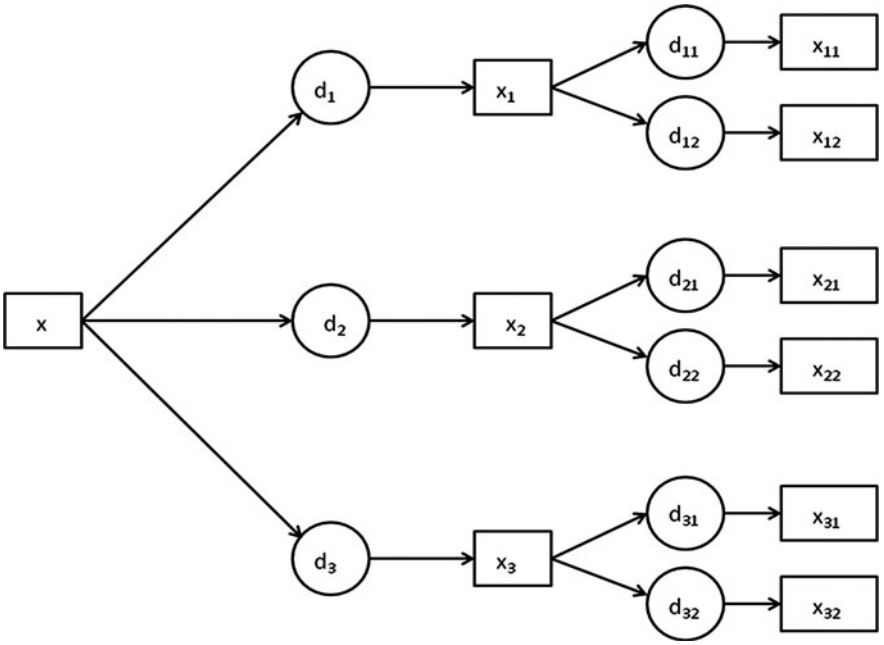


Fig. 8.1 A scenario tree for a stochastic program

8.2.2 Background and Related Work

There has been substantial effort in adding explicit SP capabilities to modeling languages. Some examples are Bisschop [2], Brooke et al. [3], Buchanan et al. [4], Entriken [7], Fourer and Lopes [9], Gassman and Ireland [10], Infanger [15], Kall and Mayer [16, 17], Kristjansson [19], Messina and Mitra [21], and Valente et al. [24].

With regard to the general theory of SP, there is an extensive literature. Birge and Louveaux [1] give a good introduction to all aspects of SP. Ruszczynski and Shapiro [23] contain a collection of 10 chapters by various SP experts on theoretical underpinnings of SP.

8.3 Steps in Building an SP Model

The approach we have taken in both LINGO and What's*Best*! for constructing an SP model is based on the following steps.

(1) Write a standard deterministic model (the core model) as if the random variables were constants.

(2) Identify the random variables, and decision variables, and their staging, i.e., the sequence in which random events occur and decisions are made.
(3) Specify the distributions describing the random variables,
(4) Specify manner of sampling from the distributions (mainly the sample size) and by implication the scenario tree,
(5) Optionally, list the variables for which we want a scenario-by-scenario report and the variables for which we want a histogram.

We illustrate above steps in both the LINGO modeling language and in the What's*Best*! spreadsheet modeling system. Our first example will be perhaps the simplest SP model possible, the newsvendor model for deciding how much to stock in advance of uncertain demand.

8.3.1 Statement/Formulation of an SP Model in LINGO

We illustrate first in LINGO.

```
! LINGO model of Newsvendor as a stochastic program.
 DATA:
   C = 30;  ! Purchase cost/unit;
   H = 5;   ! Holding cost/unit on surplus;
   P = 20;  ! Penalty cost/unit unmet demand;
   R = 65;  ! Revenue/unit sold;
   MU = 80; ! Mean demand;
   SD = 20; ! Standard deviation in demand;
 ENDDATA

! Step 1: Core model ------------------------------------+;
   MAX = PROFIT;
   PROFIT = R * SALES - C * Q - DISPOSAL_COSTS - SHORTAGE_COSTS;
   SALES + SHORT = DEMAND;
   SALES + SURPLUS = Q;
   DISPOSAL_COSTS = H * SURPLUS;
   SHORTAGE_COSTS = P * SHORT;
   @FREE(PROFIT); @FREE(SALES);

! SP related declarations -------------------------------+;
! Step 2: Stage information;
 ! Q = stage 0 decision of how much to stock;
   @SPSTGVAR( 0, Q);
 ! Demand is a random variable observed (at beginning) in stage 1;

   @SPSTGRNDV(1, DEMAND);
 ! 3) Distribution information;
   @SPDISTNORM(MU, SD, DEMAND);

 ! 4) Sample size information;
   @SPSAMPSIZE( 1, 1000);
```

Giving a guided tour of the model, the DATA section, as advised by Geoffrion in various papers, separates the data for a specific application instance from the general model equations. The Core model set of statements describe the objective and give

two equations that relate lost sales (SHORT) and left-over inventory (SURPLUS) to the amount stocked (Q) and the actual demand (DEMAND).

In step 2 we use the two qualifier functions, @SPSTGVAR(*stage*, *decision_variable*) and @SPSTRNDV(*stage*, *random_variable*) to tell LINGO that the decision variable Q must be chosen in stage 0 before the demand random variable is observed at the beginning of stage 1. An interesting feature of LINGO is that the user does not have to specify the stage of every variable. LINGO automatically infers the appropriate stage for variables for which a stage is not specified.

In step 3 the qualifier function @SPDISTNORM(MU, SD, DEMAND) tells LINGO that the random variable DEMAND has a normal distribution with mean MU and standard deviation SD. Step 4 tells LINGO to use 1000 scenarios or samples in stage 1.

Later we will describe and discuss the solution results for this model. For now we simply mention that the solution recommends setting $Q = 85.6466$ and to expect a profit of 2109.68. This newsvendor model is simple enough to be solved analytically. The analytical or "true" solution says that Q should be 85.6443 and the expected profit is 2109.94. Later we will discuss why the results based on optimizing over a sample of size 1000 are so close to the analytical solution.

8.3.2 Statement/Formulation of an SP Model in the What'sBest! Spreadsheet System

We next illustrate the same model in What's*Best*! The model specification is very similar to that in LINGO in that "qualifier" functions are used in steps 2, 3, and 4 to provide the SP-specific information. All information about the SP features is stored explicitly/openly on the spreadsheet, so that using standard Excel navigation or viewing of cells allows one to observe the SP features of the model. Figure 8.2 illustrates.

Providing a guided tour of the model, in (1) the core model is a regular, valid deterministic What's*Best*! model. You may plug in real numbers in a random cell to check results. (2) Staging information is stored about decisions in cells with qualifier functions of the form WBSP_VAR(*stage*, *cell_list*). A cell is identified as a random cell of a specified stage with a qualifier function of the form: WBSP_RAND(*stage*, *cell_list*). (3) Distribution specification is stored in a cell with the qualifier function like WBSP_DIST_NORMAL(*mean*, *standard_deviation*). (4) Sample size or number of scenarios for each stage is stored in a qualifier function of the form: WBSP_STSC(*table*). (5) Cells to be reported are listed in a qualifier function of the form WBSP_REP(*cell_list*). A cell for which we want a histogram is specified in a function of the form: WBSP_HIST(*number_bins*, *cell*);

It is possible to produce a large amount of information from an SP solution. Information on the cells listed in the WBSP_REP(*cell_list*) specification is sent to a separate tab of the worksheet as shown in Figure 8.3.

At the top of Figure 8.2 we see some summary information. In particular, the expected value for the profit is estimated to be 2109.68. We willpostpone until

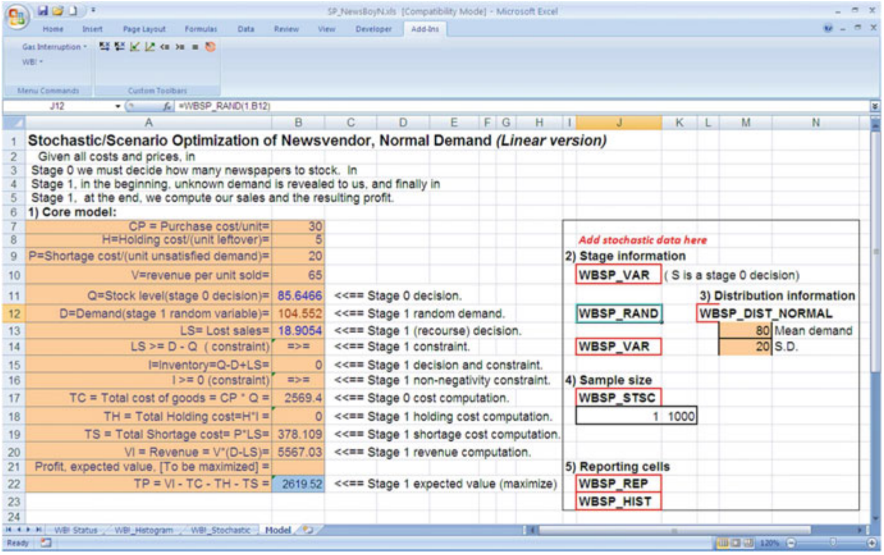


Fig. 8.2 A newsvendor model specified in What'sBest!

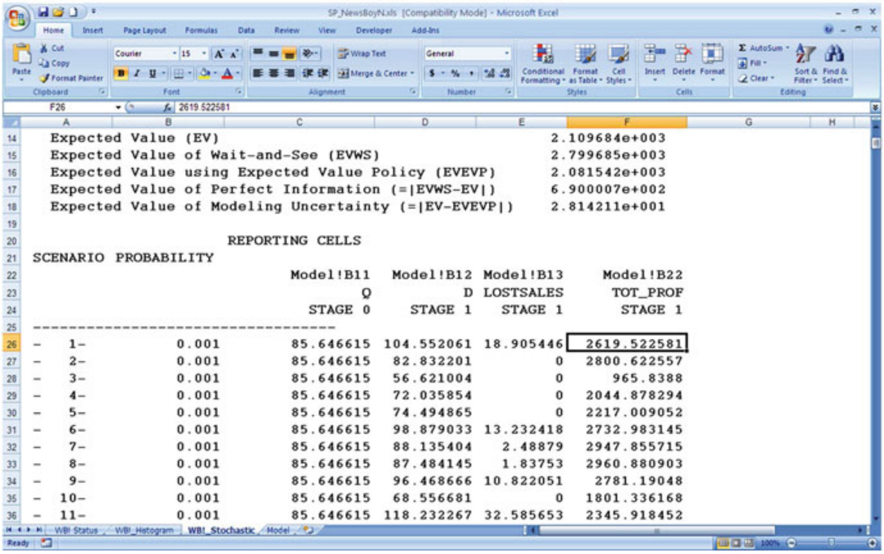


Fig. 8.3 Solution information generated from the Newsvendor model

later a discussion of the other "Expected Value" lines. One line is generated for each scenario. In Figure 8.2 we see that scenario 1 had a probability of 0.001, the amount stocked was about 85.65 (which must be the same in all scenarios), the Demand D was 104.552. This resulted in a lost sales of 18.905 and a total profit of 2619.52. Any kind of statistical analysis, such as computing standard deviations or higher moments, can be performed on the scenario data using standard Excel tools.

8.3.3 Multi-stage Models

The previous example was a model with two stages. Multiple stage models with more than two stages can be formulated and solved in the similar fashion. Below we show the formulation in LINGO of the well-known three-stage college planning model of Birge and Louveaux [1]. Some of the key things to note are the use of SPSTGVAR and SPSTGRNDV qualifier functions to specify the stages of the investment decisions and the random return outcomes.

```
! Three stage financial portfolio model. Ref. Birge & Louveaux;
! Step 1: Core Model in LINGO-------------------------------+;
SETS:
   TIME; ! Set of time periods/stages;
   ASSETS; ! Assets to invest in;
   TXA( TIME, ASSETS): RETURN, INVEST;

   SCENARIO; ! Set of possible outcomes each period;
             ! Combinations of outcomes & assets;
   SXA( SCENARIO, ASSETS): S_RETURN;
ENDSETS
 ! Decision variables...
    INVEST(t,a) = amount to invest in asset a, at end of period t.
   Random variables...
    RETURN(t,a) = growth factor for asset a, observed beginning
      period t;
DATA:
   INITIAL = 55; ! Start with $55K;
   GOAL = 80; ! Want to get at least $80K after 3 period;
   PENALTY = 4; ! Penalty for falling short of goal;
   TIME = T0..T3;
   ASSETS = BONDS, STOCKS; ! Investments available;

   SCENARIO = BONDSHI, STOCKSHI; ! Two scenarios;
     ! Outcomes for BONDS & STOCKS in each scenario;
   S_RETURN = 1.14     1.25
              1.12     1.06 ;
ENDDATA

! The core model;
! Maximize overage minus penalty for under target;
MAX = OVER - PENALTY * UNDER;

! Initial allocation;
[R_INIT] @SUM( ASSETS( A): INVEST( 1, A)) = INITIAL;

! Portfolio value in period t;
@FOR( TIME( T) | T #GT# 1:
  @SUM( ASSETS( A): INVEST( T, A)) =
  @SUM( ASSETS( A): RETURN( T, A) * INVEST( T - 1, A));
    );
FINAL = @SUM( ASSETS( A): INVEST( @SIZE( TIME), A));
OVER - UNDER = FINAL - GOAL ;

! SP Related Declarations ----------------------------+;
! Step 2) Stage information;
!    Declare the stage of each decision variable;
```

```
@FOR( TXA( T, A):
   @SPSTGVAR( T-1, INVEST( T, A));
    );
! The stages of the return random variables;
@FOR( TXA( T, A) | T #GT# 1:
   @SPSTGRNDV( T-1, RETURN( T, A));
    );
! Step 3, the distributions;
! Construct a discrete distribution table, D1;
! Declare a discrete distribution table D1;
@SPTABLESHAPE( 'D1', @SIZE( SCENARIO), @SIZE( ASSETS));
! Fill the distribution D1,...;
@FOR( SCENARIO( s):
  @SPTABLEOUTC( 'D1', 1/@SIZE( SCENARIO)); ! Probabilities 1st;
  ! and then the actual outcomes;
  @FOR( ASSETS( A): @SPTABLEOUTC( 'D1', S_RETURN( s, A)));
    );
  ! Now specify that each stage has the same distribution D1;
@FOR( TIME( T) | T #GT# 1:
   ! Declare an instance of our parent distribution;
   @SPTABLEINST( 'D1', TIME( T));
   ! Bind the random variables to the instance;
   @FOR( ASSETS( A):
      @SPTABLERNDV( TIME( T), RETURN( T, A))
        );
    );
```

The same three-stage model in What's*Best*! is shown in Figure 8.4.

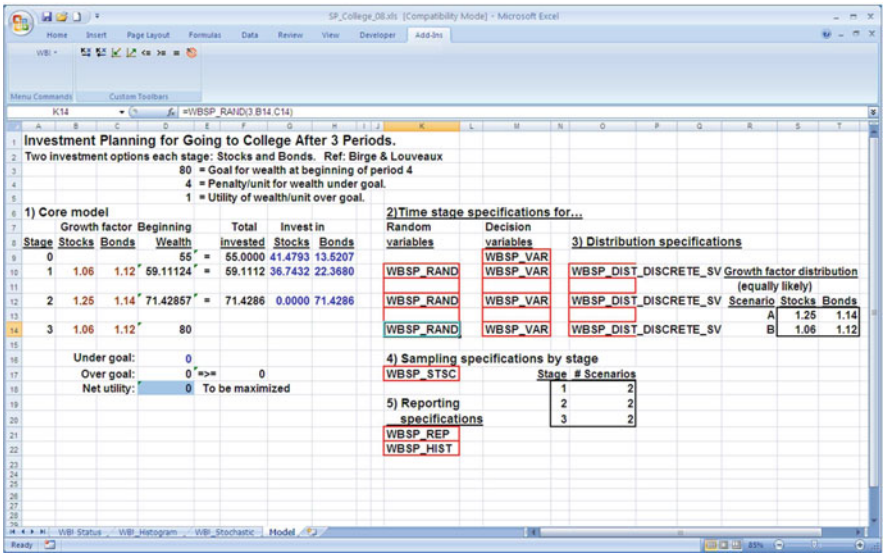
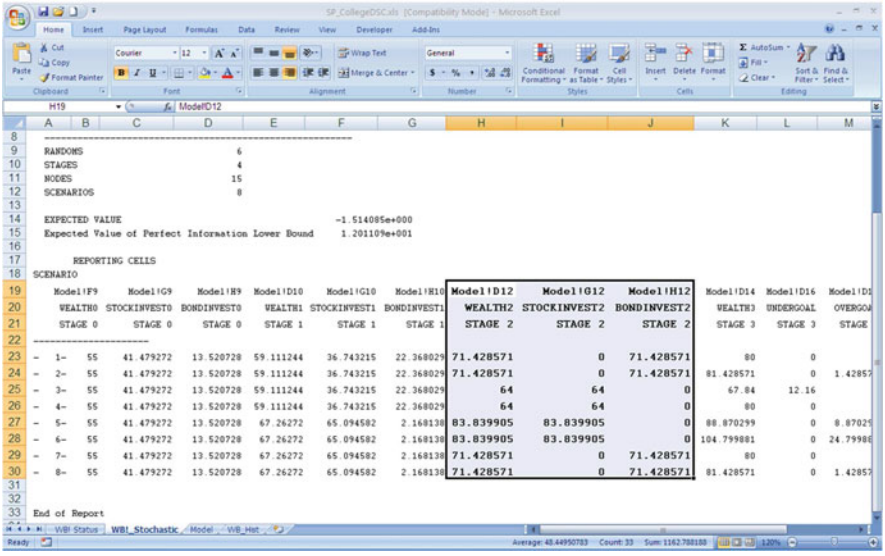


Fig. 8.4 Optimal portfolio reinvestment over three periods

The essential formulae of the core model are in column D where the beginning wealth in stage t is set equal to the amount invested in each of stocks and bonds, columns G and H in the preceding stage, times the (random) growth factors in columns B and C, e.g., D12 = SUMPRODUCT(G10:H10,B12:C12). The over- or underachievement of goal is computed with D17 = D14 − D3 + D16. Cell E17 constrains D17 >= 0. Columns K and M specify the stages for the decision variables, columns G and H, and the random cells, columns B and C. The distribution of two possible outcomes each period is specified in the cell range O10:T14. Thus, there are two possible stage scenarios in each stage, see cells K16:O20. Cell K21 asks for a scenario-by-scenario report of certain cells with the qualifier: WBSP_REP(F9,G9,H9,D10,G10,H10,D12,G12,H12,D14,D16,D17,D18).

Two possible outcomes in each of three periods mean $2^3 = 8$ full scenarios in total. This scenario-by-scenario report appears in the Excel tab displayed in Figure 8.5. Notice the interesting behavior of the optimal policy in stage 2. We want to maximize the wealth at the end of stage 3; however, there is a heavy penalty (of 4) for falling short of the target of 80. Notice that if the beginning wealth is either very low (64) or very high (83.8399), we invest everything in STOCK, the investment with the higher expected return, even though it is the riskier one. The reasoning is if we are at 64, we know we will fall short of the goal, so we might as well minimize the expected amount short. If we are at 83.8399, we know we will achieve our goal, so we might as well maximize the expected amount by which we exceed our goal.

If we are at an intermediate level (71.428571), we invest everything in bonds because it will safely guarantee that we will just achieve our goal, regardless of which two scenarios next occur.



	A	B	C	D	E	F	G	H	I	J	K	L	M
9	RANDOMS			6									
10	STAGES			4									
11	NODES			15									
12	SCENARIOS			8									
14	EXPECTED VALUE					-1.514085e+000							
15	Expected Value of Perfect Information Lower Bound					1.201109e+001							
17	REPORTING CELLS												
18	SCENARIO												
19		Model!F9	Model!G9	Model!H9	Model!D10	Model!G10	Model!H10	Model!D12	Model!G12	Model!H12	Model!D14	Model!D16	Model!D1
20		WEALTH0	STOCKINVEST0	BONDINVEST0	WEALTH1	STOCKINVEST1	BONDINVEST1	WEALTH2	STOCKINVEST2	BONDINVEST2	WEALTH3	UNDERGOAL	OVERGOA
21		STAGE 0	STAGE 0	STAGE 0	STAGE 1	STAGE 1	STAGE 1	STAGE 2	STAGE 2	STAGE 2	STAGE 3	STAGE 3	STAGE
23	- 1-	55	41.479272	13.520728	59.111244	36.743215	22.368029	71.428571	0	71.428571	80	0	
24	- 2-	55	41.479272	13.520728	59.111244	36.743215	22.368029	71.428571	0	71.428571	81.428571	0	1.42857
25	- 3-	55	41.479272	13.520728	59.111244	36.743215	22.368029	64	64	0	67.84	12.16	
26	- 4-	55	41.479272	13.520728	59.111244	36.743215	22.368029	64	64	0	80	0	
27	- 5-	55	41.479272	13.520728	67.26272	65.094582	2.168138	83.839905	83.839905	0	88.870299	0	8.87029
28	- 6-	55	41.479272	13.520728	67.26272	65.094582	2.168138	83.839905	83.839905	0	104.799881	0	24.79988
29	- 7-	55	41.479272	13.520728	67.26272	65.094582	2.168138	71.428571	0	71.428571	80	0	
30	- 8-	55	41.479272	13.520728	67.26272	65.094582	2.168138	71.428571	0	71.428571	81.428571	0	1.42857
33	End of Report												

Fig. 8.5 Optimal policy for portfolio reinvestment over three periods

8.4 Scenario Generation

A crucial capability of an SP modeling system is that of populating the scenario tree with appropriate random values. We break this down into three steps: (1) generating uniform random numbers, (2) converting uniform random numbers into random numbers of a specified general distribution, and (3) inducing correlation between two or more random variables, or more generally, generating a vector of random variables with an appropriate joint distribution. In terms of design, all aspects of scenario generation are contained in random variable generation component of the LINDO API. The user front end, LINGO or What's*Best*! in our case, need not be concerned with scenario generation other than getting from the user the stage information, the distribution information, and the sampling choices.

8.4.1 Uniform Random Number Generation

An important component of any Monte Carlo package is a pseudo-uniform random number generator. The LINDO API has three options for generating uniforms: (1) the classic 31 bit linear congruential generator, (2) a composite linear congruential generator, and (3) a Mersenne twister generator. The default generator is (2), the composite generator, see L'Ecuyer et al. [20]. The stream of uniforms in $(0,1)$, $u[n]$ is generated by the recursion:

$$\begin{aligned}
x[n] &= (1403580 \cdot x[n-2] - 810728 \cdot x[n-3]) \mod 4294967087;\\
y[n] &= (527612 \cdot y[n-1] - 1370589 \cdot y[n-3]) \mod 4294944443;\\
z[n] &= (x[n] - y[n]) \mod 4294967087;\\
u[n] &= z[n]/4294967088 \quad \text{if } z[n] > 0;\\
&= 4294967087/4294967088 \quad \text{if } z[n] = 0.
\end{aligned}$$

This generator has cycle length of about $2^{191} = 3.14 \cdot 10^{57}$. This is a considerable improvement over $2^{31} = 2.15 \cdot 10^9$ cycle length for tradition single stream 31-bit generators. It has good multidimensional uniformity up to about 45-dimensional hypercubes.

8.4.2 Random Numbers from Arbitrary Distributions

The LINDO API is able to generate random variables from about two dozen standard distributions, including Beta, Binomial, Cauchy, Chisquare, Exponential, F, Gamma, Geometric, Gumbel, Hypergeometric, Laplace, Logarithmic, Logistic,

Lognormal, Negative binomial, Normal, Pareto, Poisson, Student t, Triangular, Uniform,Weibull. In addition jointly distributed random variables can be generated from a user specified Discrete/Empirical/Joint table of outcomes.

All general random variables are generated by the inverse transform method. Suppose a random variable has a cumulative distribution function (cdf), $F(x) = \text{Prob}\{\text{the random variable} \leq x\}$. The basic steps are as follows:

(1) Generate a uniform random number, u, in $(0, 1)$.
(2) Convert the uniform to the desired distribution by inverting the cdf, that is, invert the function $u = F(x)$ to solve for x in terms of u: $x = F^{-1}(u)$. A graph, see Figure 8.6, perhaps explains it best.

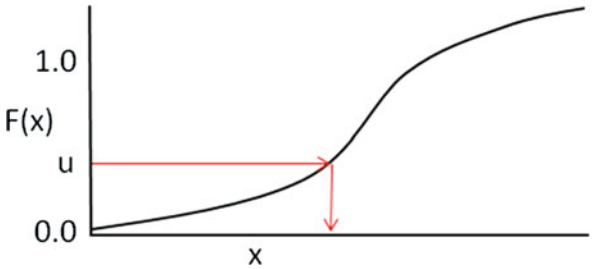


Fig. 8.6 Inverse transform method, graphically

8.4.3 Quasi-random Numbers and Latin Hypercube Sampling

The default sampling method in LINDO API is Latin hypercube sampling. If you ask it to generate 100 random numbers uniformly distributed in $(0, 1)$, it will (1) divide the interval $(0, 1)$ into 100 equal subintervals and (2) generate one random number uniformly distributed over each subinterval. This method has an important qualitative feature and an important theoretical feature, namely (a) the distribution appears more uniform than if one had taken a purely random sample and (b) nevertheless, it is unbiased in that every point in $(0, 1)$ has equal probability of being chosen. The inverse transform method works nicely with Latin hypercube sampling. If we generate a sample of 100 normal random variables using this combination of methods, the sample has the nice feature that each percentile of the normal distribution will have one sample drawn from it. Figures 8.7 and 8.8 illustrate this feature. We took a sample of 100 from a normal distribution with mean 100 and standard deviation 10. Notice that the Latin hypercube sample not only looks more normal, but the sample mean and standard deviation more closely approximate the population mean.

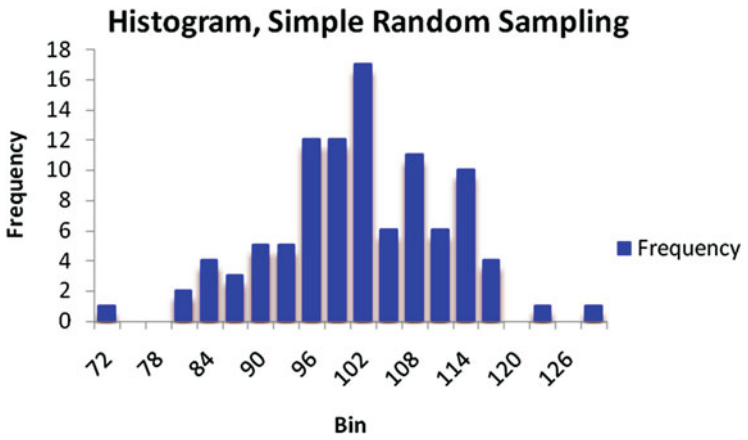


Fig. 8.7 Sample of 100 normal deviates using pure random sampling; mean = 100.31, sd = 10.14

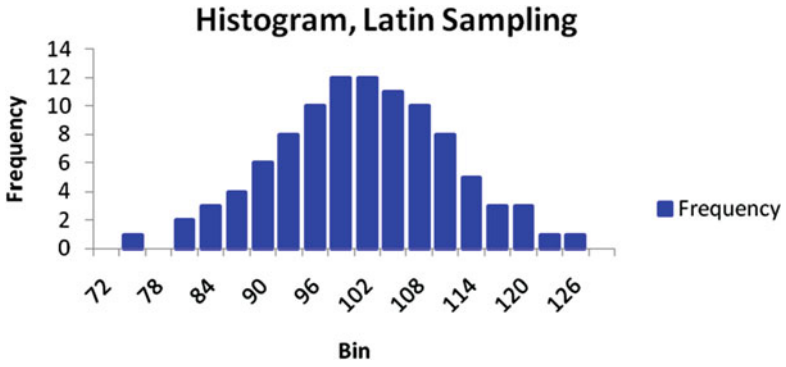


Fig. 8.8 Sample of 100 normal deviates using Latin hypercube sampling; mean = 99.98, sd = 9.98

8.4.4 Generating Correlated Random Variables

There are three traditional ways of measuring correlation between two random variables x and y, the traditional Pearson "linear" correlation taught in Stat 101 and two rank correlation methods, Spearman rank and Kendall tau rank. The LINDO API allows the user to choose which of the three is to be used in generating correlated random variables. Pearson correlation makes sense for normal random variables. For arbitrary distributions, the two rank correlation measures may be more convenient. All three are summarized below.

Pearson

Define

$$\bar{x} = \sum_{i=1}^{n} x_i/n; \qquad s_x = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)},$$

then the Pearson correlation is defined as

$$\rho_s = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n s_x s_y).$$

Spearman Rank

Same as Pearson, except x_i and y_i are replaced by their ranks, with minor adjustments if there are ties, e.g., if $x_i = x_{i+1}$.

Kendall Tau Rank

Here

$$\rho_\tau = \sum_{i=1}^{n} \sum_{k=i+1}^{n} 2\,\text{sign}[(x_i - x_k)(y_i - y_k)]/[n(n-1)],$$

where $\text{sign}(x) = -1$, 0, or $+1$ depending on whether $x < 0$, $x = 0$, or $x > 0$.

The Kendall correlation has a simple probabilistic interpretation. If (x_1, y_1) and (x_2, y_2) are two observations on two random variables that have a Kendall correlation of ρ_k, then the probability that the two random variables move in the same direction is $(1+ \rho_k)/2$. That is,

$$\text{Prob}\{(x_2 - x_1)(y_2 - y_1) > 0\} = (1 + \rho_k)/2.$$

For example, if the weekly change in the DJI and the SP500 has a Kendall correlation of 0.8, then the probability that these two indices will change in the same direction next week is $(1+0.8)/2 = 0.9$. Although the Kendall rank correlation has this simple interpretation, the Spearman correlation contains more information. For example, if you have a sample of size four for two random variables, there are only seven possible values, -1, $-2/3$, $-1/3$, 0, $1/3$, $2/3$, 1, for the Kendall correlation, whereas there are 11 possible values for the Spearman correlation.

A useful feature of rank correlation is that it is unchanged by a monotonic increasing transformation, such as the inverse transform method. Thus, if we can generate two uniform random variables u and v with a certain rank correlation and we generate two random variables x and y from arbitrary distributions by the inverse transform method, i.e., $x = F_x^{-1}(u)$ and $y = F_y^{-1}(v)$, then x and y will have the same rank correlation as u and v.

8.5 Solution Output for an SP Model

In the process of solving an SP model, a lot of information is generated. How is this information best summarized and presented to the user?

8.5.1 Histograms

One advantage of SP, as well as simulation, relative to an analytic solution of a model is that a good portrayal of the distribution of various outcomes such as profit is available. Sometimes the distribution of an outcome random variable may be surprising. In What's*Best*!, one can request a histogram with 15 bins of cell TOT_PROF by inserting the qualifier = WBSP_HIST(15,TOT_PROF) somewhere in the sheet. As an example, consider the standard newsboy problem with normal distributed demand. One might expect that if demand is normal distributed, then profit might also be approximately normal distributed. The histogram in Figure 8.9, based on our earlier newsboy example, shows that such is definitely not the case.

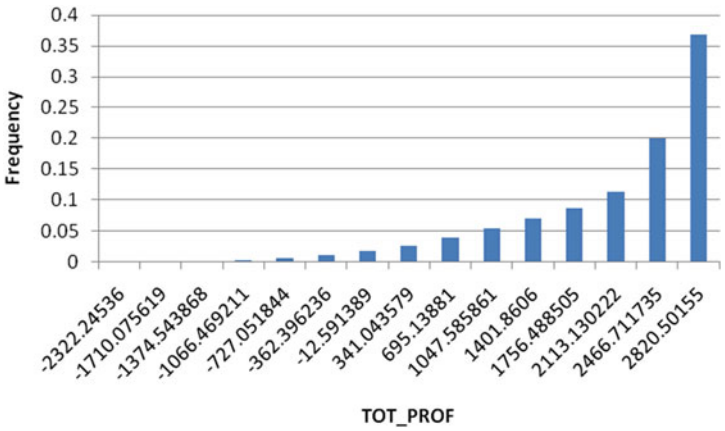


Fig. 8.9 Empirical distribution of profit from newsvendor optimal policy

Another example is a put option, i.e., the ability to sell a share of a specific stock at a specified strike price at some point in the future. You do not want to exercise the option in the current period if the current price is above the strike price. If the current price is below the strike price, you may want to exercise, but you may also want to consider waiting in case later the price drops even lower. SP can be used to find an optimal exercise policy. An important question is the expected value of such an option, assuming an optimal exercise policy is followed. For a certain such option the optimal policy was found by SP and its expected present value was determined

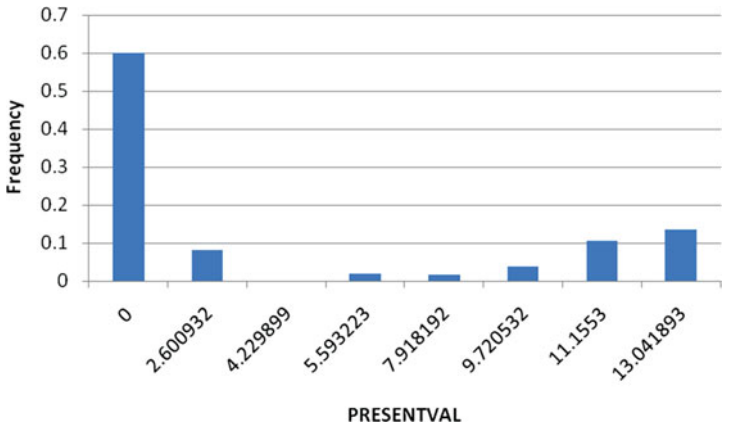


Fig. 8.10 Empirical distribution of present value of a put option under optimal policy

to be $3.845. One might expect that the typical value of the option would be about $3.85. The graph in Figure 8.10 shows that, in fact, outcomes near $3.85 are rather unlikely. About 60% of the time the option expires, unused, and is worthless. About 20% of the time the option is worth around $12.

Making histograms available to the user presents an interesting design question: where should histogram construction be implemented, in the front end or in the solver engine? We have implemented histogram construction in the solver engine, LINDO API, but leave display of the histogram to the front end. The user has the option of specifying the number of bins in advance. If the number of bins is not specified, then the API chooses the number of bins based on aesthetic considerations and the number of scenarios. If a histogram has too few bins, then the histogram will look "saw-toothed" or lumpy. If the histogram has too many bins, then the histogram may look ragged or erratic because of randomness in the number of scenarios that fall in a given bin. A heuristic is used to strike a compromise between these two considerations.

8.5.2 Expected Value of Perfect Information and Modeling Uncertainty

A current user of SP is interested in two things, what is the expected profit of an optimal policy and what are the actions or decisions to be taken now in stage 0 in order achieve this profit in expectation? Both LINGO and What's*Best*! report the expected profit in solution summary information and directly display the optimal stage 0 decisions.

Before even using SP, a thoughtful user might ask the following two questions: (1) How much can I improve my expected profits by using SP and (2) How much is uncertainty costing me, e.g., if I had perfect forecasts, how much could I improve

profits? The answers to these two questions can be any one of the four combinations of "a lot" and/or "not much." For example, in a newsvendor-like inventory problem, if the cost/unit of carrying too much is about equal to the cost/unit of carrying too little, then the value of using SP is not much relative to just stocking as if demand will always be equal to the mean. On the other hand, if the variance in demand is high, then the value of having better forecasts may be a lot. There are other situations where just the reverse is true, i.e., the value of using SP is a lot, even though the value of better forecasts is not much. LINGO and What's*Best*! supply two statistics, EVMU (Expected Value of Modeling Uncertainty) and EVPI (Expected Value of Perfect Information). Slightly more explicitly

EVPI = Expected increase in profit if we know the future in advance.
EVMU = Expected decrease in profit if we replaced each random variable by a single estimate and act as if this value is certain.

In the SP literature, EVMU is sometimes called VSS (Value of Stochastic Solution). Let us look at how EVMU and EVPI are provided in LINGO for the Newsvendor model considered previously. The solution summary section is as follows:

```
Objective (EV):                              2109.684
Wait-and-see model's objective (WS):         2799.685
Perfect information (EVPI = |EV - WS|):      690.0007
Policy based on mean outcome (EM):           2081.542
Modeling uncertainty (EVMU = |EM - EV|):     28.14211
```

The first line says that, given the problem as stated, the estimate of expected profit is 2109.684. The second, "Wait-and-see," line says that if we could postpone our inventory stocking decision until we saw demand (alternatively, we have perfect forecasts), then our expected profit is estimated to be 2799.685. The third line, EVPI, is the estimated amount of additional profit from having this information.

The fourth line, EM, reports our estimated expected profit if we acted as if the demand would always be the mean. In this case we would stock the mean, 80, rather than the SP recommended level of 85.647. Thus, we would incur higher shortage costs than under the optimal policy.

Figure 8.11 lays out graphically the relationship between doing the best job possible with the available information (EVMU) and the benefit of getting perfect information (EVPI).

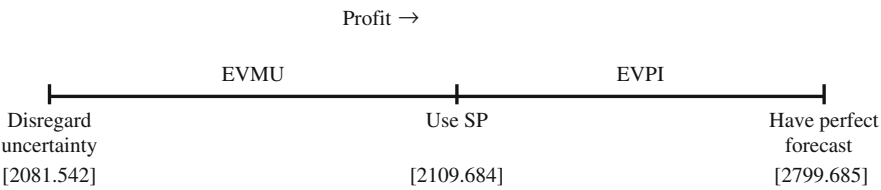


Fig. 8.11 Relationship between good decision making plus good information

The above is for a two-stage SP. EVPI generalizes easily to more than two stages; however, it is not so straightforward to generalize EVMU to more than two stages.

8.6 Conclusions

We have shown two approaches, in a modeling language and in a spreadsheet, to making SP available to typical modeling analysts. There are two challenges in making SP available to wider audience: computability and usability. Users expect to be able to solve nontrivial problems with an SP modeling capability. In this sense, SP is similar to integer programming. Unsophisticated users can easily formulate relatively simple looking models that take very long to solve. Much good effort by talented researchers has been devoted to methods for solving SP problems. Less effort has been devoted to usability, although our experience is that the challenge there is almost as great. We hope that this chapter illustrates that many of the traditional theories of good modeling and good model system design apply to SP just as well to more traditional types of operations research models.

References

1. Birge J, Louveaux F (1987) Introduction to stochastic programming. Springer, New York, NY
2. Bisschop J (2006) AIMMS-optimization modeling. Lulu Press, Haarlem, The Netherlands
3. Brooke A, Kendrick D, Meeraus A (1992) GAMS, a user's guide. The Scientific Press, Redwood City, CA
4. Buchanan C, McKinnon K, Skondras G (2001) The recursive definition of stochastic linear programming problems within an algebraic modeling language. Annals of Operations Research 104(1):15–32
5. Carino D, Ziemba W (1998) Formulation of the Russell–Yasuda Kasai financial planning model. Operations Research 46:433–449
6. Dantzig G (1963) Linear programming and extensions. Princeton University Press, Princeton, NJ
7. Entriken R (2001) Language constructs for modeling stochastic linear programs. Annals of Operations Research 104(1):49–66
8. Eppen G, Martin R, Schrage L (1989) A scenario approach to capacity planning. Operations Research 37(4):517–527
9. Fourer R, Lopes L (2009) StAMPL: A filtration-oriented modeling tool for multistage stochastic recourse problems. INFORMS Journal of Computing 21(2):242–256
10. Gassmann HI, Ireland AM (1996) On the formulation of stochastic linear programs using algebraic modeling languages. Annals of Operations Research 64:83–112
11. Gaustad G, Li P, Kirchain R (2007) Modeling methods for managing raw material compositional uncertainty in alloy production. Resources Conservation & Recycling 52:180–207
12. Geoffrion AM (1976) The purpose of mathematical programming is insight, not numbers. Interfaces 7(1):81–92
13. Geoffrion AM (1989) The formal aspects of structured modeling. Operations Research 37(1):30–51
14. Geoffrion AM (1992) Indexing in modeling languages for mathematical programming. Management Science 38(3):325–344

15. Infanger G (1999) GAMS/DECIS user's guide. http://www.gams.com/dd/docs/solvers/decis.pdf
16. Kall P, Mayer J (1996) An interactive model management system for stochastic linear programs. Mathematical Programming 75:221–240
17. Kall P, Mayer J (2005) Stochastic linear programming: Models, theory, and computation. Springer
18. Knowles T, Wirick J (1988) Peoples gas light and coke company plans gas supply. Interfaces 28(5):1–12
19. Kristjansson B (2005) MPL user manual. Maximal Software, Arlington, VA
20. L'Ecuyer P, Simard R, Chen E, Kelton W (2002) An object-oriented random-number package with many long streams and substreams. Operations Research 50(6):1073–1075
21. Messina E, Mitra G (1997) Modelling and analysis of multistage stochastic programming problems: A software environment. European Journal of Operations Research 101:343–359
22. Pereira M, Pinto L (1991) Multi-stage stochastic optimization applied to energy planning. Mathematical Programming 52(2):359–375
23. Ruszczynski A, Shapiro A (2003) Stochastic programming. Handbooks in operations research and management science, vol 10. Elsevier, Amsterdam
24. Valente C, Mitra G, Sadki M, Fourer R (2009) Extending algebraic modeling languages for stochastic programming. INFORMS Journal of Computing 21(1):107–122
25. Wallace SW, Ziemba WT (2005) Applications of stochastic programming. MPS-SIAM Series on Optimization, Philadelphia, PA

Chapter 9
Advances in Business Analytics at HP Laboratories

Business Optimization Lab, HP Labs, Hewlett-Packard

Abstract HP Labs' Business Optimization Lab is a group of researchers focused on developing innovations in business analytics that deliver value to HP. This chapter describes several activities of the Business Optimization Lab, including work in product portfolio management, prediction markets, modeling of rare events in marketing, and supply chain network design.

9.1 Introduction

Hewlett-Packard is a technology company that operates in more than 170 countries around the world. HP explores how technology and services can help people and companies address their problems and challenges and realize their possibilities, aspirations, and dreams.

HP provides infrastructure and business offerings ranging from handheld devices to some of the world's most powerful supercomputer installations. HP offers consumers a wide range of products and services from digital photography to digital entertainment and from computing to home printing. HP was founded in 1939. Its corporate headquarters are in Palo Alto, CA. HP is among the world's largest IT companies, with revenue totaling $118.36 billion for the fiscal year that ended Oct 31, 2008.

HP's three business groups drive industry leadership in core technology areas:

- Personal Systems Group: business and consumer PCs, mobile computing devices and workstations.

Dirk Beyer, M-Factor, Inc. • Scott Clearwater • Kay-Yut Chen, HP Labs • Qi Feng, McCombs School of Business, University of Texas at Austin • Bernardo A. Huberman, HP Labs • Shailendra Jain, HP Labs • Zainab Jamal, HP Labs • Alper Sen, Department of Industrial Engineering, Bilkent University • Hsiu-Khuern Tang, Intuit • Bob Tarjan, HP Labs • Krishna Venkatraman, Intuit • Julie Ward, HP Labs • Alex Zhang, HP Labs • Bin Zhang, HP Labs

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_9,

- Imaging and Printing Group: Inkjet, LaserJet and commercial printing, printing supplies, digital photography and entertainment.
- Enterprise Business Group: enterprise services, business products including storage and servers, software and technology services for customer support.

At its heart, HP is a technology company, fueled by progress and innovation. The majority of HP's research is conducted in our business groups, which develop the products and services we offer to customers. As Hewlett-Packard's central research organization, HP Labs' role is to invent for the company's future.

HP Labs' function is to deliver breakthrough technologies and technology advancements that provide a competitive advantage for HP and to create business opportunities that go beyond HP's current strategies. The lab also helps shape HP strategy, and it invests in fundamental science and technology in areas of interest to HP.

For more than 40 years, HP Labs has been advancing technology and improving the way our customers live and work. From the invention of the desktop scientific calculator and the HP LaserJet printer to blade technology innovations and power-efficiency improvements for data centers, HP Labs is continuously pushing the boundaries of research to deliver more valuable technology experiences.

With 600 researchers across 23 labs in seven worldwide locations, HP Labs brings together some of the most distinguished researchers across a diverse set of scientific and technical disciplines—including experts in economics, science, physics, computer science, sociology, psychology, mathematics, and engineering.

These dedicated researchers are tackling some of the most important challenges of the next decade through a focus on high-impact research, a commitment to open innovation, and a drive to transfer technology to the marketplace. HP Labs' goal is to create breakthrough technology experiences for individuals and businesses around the world.

HP's deep roots in technologies and very competitive business environment provide a very rich set of opportunities for applied research in advanced analytics. Some of this applied research thrust in analytics is directed toward new product or service creations, though the major share of activities is geared toward operational processes innovation. This chapter describes selected activities of HP Labs' Business Optimization Lab, a group focused on advancing technologies and building high-impact innovative applications for operations and personalization, both driven by advanced analytics.

The researchers in the Business Optimization Lab exploit opportunities to build upon existing methodologies and create advanced analytics models and solutions for a comprehensive array of business contexts. The applications of this work span a wide range of areas including marketing, supply chain management, enterprise-wide risk management, service operations, and new service creation. Methodologies driving this applied research at HP Labs include operations research, industrial engineering, economics, statistics, marketing science, and computer science. For a summary of these activities see Jain [15].

9.1.1 Diverse Applied Research Areas with High Business Impact

This chapter presents four applied research projects conducted in the Business Optimization Lab that address HP's business needs in diverse areas.

The first study describes HP Labs' work in product variety management, which is at the interface of marketing and supply chain management decisions. Conventional wisdom suggests that a manufacturer should offer a broad variety of products in order to meet the needs of a diverse set of customers. While this is true to an extent, product variety comes with significant operational costs, which in excess may be counter-productive to profitability. Since the 1990s HP has faced many of these challenges due to its vast product portfolio. Business units sought methods to understand the costs of complexity and to identify which products were truly important to their business, so that they could refine their product offering without compromising revenue. To address these challenges, HP Labs introduced a new metric, coverage, for evaluating product portfolios in configurable product businesses. Coverage looks beyond the individual performance of products and considers their interdependence through orders. This metric, and HP Labs' accompanying Revenue Coverage Optimization tool (RCO), enables HP to identify products most critical to its offering, as well as candidates for discontinuance. As a result, HP has improved its operational focus on key products while also reducing the complexity of its product offering, leading to significant business benefits.

The second section describes the methodology and application of prediction market for forecasting business events, when markets are not efficient. Forecasting has been important since the dawn of business. There are two approaches in the context of using information for forecasting. The popular approach, backed up by decades of development of computing technologies, is the use of statistical analysis on historical data. This approach can be very successful when the relevant information is captured in historical data. In many situations, however, there is either no historical data or the data contain no patterns useful for forecasting. A good example is forecasting the demand of a new product. Thus, a second approach is to tap into tacit and subjective information in the minds of individuals. This so-called wisdom of crowds phenomenon has been documented over the centuries. The prediction markets, where people are allowed to interact in organized markets governed by well-defined interaction rules, have been shown to be an effective way to tap into the collective intelligence of crowds. If these markets are large enough and properly designed, they can be more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Forecasting business events, on the other hand, may involve only a handful of busy experts, and they do not constitute an efficient market. We describe an alternate method of harnessing the distributed knowledge of a small group of individuals by using a two-stage mechanism. This mechanism is designed to work on small groups, or even an individual. This technique has been applied to several real-world demand forecasting problems. We will present a case study of its use in demand forecasting a technology hardware product and also discuss issues about real-world implementation.

In the third area, we describe modeling of rare events in marketing. A rare event is an event with a very small probability of occurrence. Typical examples of such events from social sciences that readily come to mind are wars, outbreak of infections, and breakdown of a city's transport system or levies. Examples of such events from marketing are in the area of database marketing (e.g., catalogs, newspaper inserts, direct mailers sent to a large population of prospective customers) where only a small fraction (less than 1%) responded resulting in a very small probability of a response (event). More recent examples of rare events have emerged in marketing with the advent of the Internet and digital age and the use of new types of marketing instruments. A firm can reach a large population of potential customers through its web site, display ads, e-mails, and search marketing. But only a very small proportion of those exposed to these instruments respond. To make business and policy planning more effective it is important to be able to analyze and predict these events accurately. Rare event variables have been shown to be difficult to predict and analyze. There are two sources of the problem. The first source is that standard statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events. The second source of the problem is that commonly used data collection strategies are grossly inefficient for rare events data. In this study we share a choice-based sampling approach to discrete-choice models and decision-tree algorithms to estimate the response probabilities at the customer level to a direct mail campaign when the campaign sizes are very large (in millions) and the response rates are extremely low. We use the predicted response probabilities to rank the customers which will allow the business to run targeted campaigns.

In our fourth and last study, we describe a mathematical programming model that constitutes the core of a number of analytical decision support applications for decision problems ranging from design of manufacturing and distribution networks to evaluation of complex supplier offers in logistics procurement processes. We provide some details on two applications of the model to evaluate various distribution strategy alternatives. In these applications, the model helps answer questions such as whether it is efficient to add more distribution centers to the existing network and which distribution centers and transport modes are to be used to supply each customer location and segment, by quantifying the trade-off between the supply chain costs and order cycle times.

9.2 Revenue Coverage Optimization: A New Approach for Product Variety Management

HP's Personal Systems Group (PSG) is a $40B business that sells workstations, desktops, notebooks, and handheld devices to consumers and businesses. In October 2004, PSG offered tens of thousands of distinct products in its product lines. PSG's Global Business Unit Team knew their large and complex product offering led to confusion among sales people and customers, high administrative costs for forecasting and managing inventory of each product, and, most seriously,

poor order cycle time (OCT). A typical PSG order consists of many products, and an order does not ship until each of its products is available, so a stock-out of a single product delays the entire order. Because PSG's product line was so large, it was difficult and costly to maintain adequate availability for all products. Consequently, PSG's average OCT ranged from 11 to 14 days in North America (depending on the product line) compared to 5–7 days for the leading competitor. This difference adversely affected HP's customer satisfaction and market share.

The PSG team sought to identify a "Core Portfolio" of products that were most important to achieve their business goals. Once these Core products were identified, PSG could reduce the wait time for these products by renegotiating supply contracts and increasing inventory as needed. PSG also hoped to identify lower-priority products and either eliminate them from the product offering or offer them with longer lead times than Core Portfolio products. Prior to 2004, PSG used revenue thresholds as the measure for product importance. However, revenue is an insufficient criterion because it fails to recognize that some low-revenue products, such as power supplies, are critical to fulfilling many orders. PSG needed a more effective way to measure each product's importance.

Similar product proliferation issues affected other parts of HP, including Business Critical Systems (BCS). Business leaders sought the help of OR researchers and practitioners in the company to manage HP's product portfolio in a disciplined manner. As a result, HP created two powerful OR-based solutions for managing product variety (see Ward et al. [29].) The first solution, developed by HP's Strategic Planning and Modeling (SPaM) group, is a framework for screening new products *prior to introduction*. It uses custom-built return-on-investment (ROI) calculators to evaluate each proposed new product; those that do not meet a threshold ROI level are targeted for exclusion from the proposed lineup. The second, HP Labs' Revenue Coverage Optimization (RCO) tool, is used to manage product variety *after introduction*. RCO enables HP businesses to increase operational focus on their most critical products. Together, these tools have enabled HP to streamline its product offerings, improve execution, achieve faster delivery, lower overhead, and increase customer satisfaction and market share.

This chapter focuses on the second solution. It describes the RCO technology for managing product variety after it has been introduced into the portfolio and its implementation in HP. The next section introduces the metric of coverage for evaluating a product portfolio and describes the evolution of approaches that led to a fast new maximum flow algorithm for revenue coverage optimization. The subsequent sections present the results achieved through the use of RCO in HP, followed by concluding remarks.

9.2.1 Solution

9.2.1.1 Coverage: A New Metric for Product Portfolios

The joint business unit and HP Labs team knew that when determining the importance of products in an existing product portfolio, it would not suffice to examine

each product in isolation in order history, particularly in a business where orders consist of many interdependent items. As mentioned previously, a product that generated relatively little revenue of its own could, in fact, be a critical component to some large-revenue orders, and therefore be essential to order fulfillment. To address this, HP Labs developed a new metric of a product portfolio that captures the interrelationship among products and orders. This metric, called *order coverage*, represents the percentage of past orders that could be completely fulfilled from the portfolio. Similarly, *revenue coverage* of a portfolio is the revenue of its covered orders as a percentage of the total revenue of orders in the data set. The concept of coverage provides a meaningful way of measuring the overall impact of each product on a business. The tool we developed, called the Revenue Coverage Optimization (RCO) Tool, finds the smallest portfolio of products that covers any given percentage of historical order revenue.[1] More generally, given a set of historical orders, RCO computes a nested series of product portfolios along the efficient frontier of order revenue coverage and portfolio size.

The black curve in Figure 9.1 illustrates this efficient frontier. In this example, 80% of order revenue can be covered with less than 27% of the total product portfolio, if those products are selected according to RCO's recommendations. One can use this tool to select the portfolio along the efficient frontier that offers the best trade-off—relative to their business objectives—between revenue coverage and portfolio size. The strong Pareto effect in the RCO curve presents an important

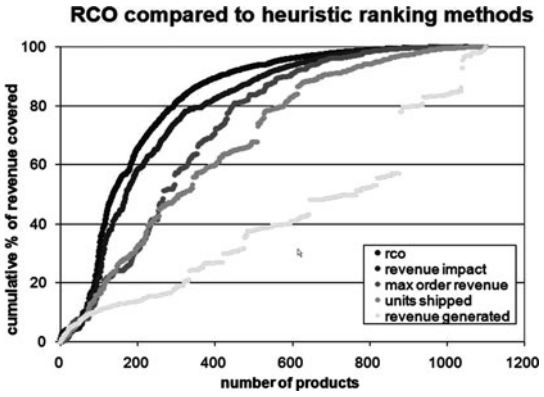


Fig. 9.1 This chart shows revenue coverage vs. portfolio size achieved by RCO (*black*) and four other product ranking methods, applied to the same historical data. The four other curves, in decreasingly saturated grays, are based on ranking by the following product metrics: revenue impact (the total revenue of orders containing the product); maximum revenue of orders containing the product; number of units shipped; and finally, individual product revenue

[1] In a nutshell, the RCO tool answers questions like "If I can pick only 100 products, which ones should I choose so I can maximize the revenue from orders that *only* have these products in it?" We argue, this is a better question to ask than "Which 100 products sold the most units?" or "Which 100 products show the highest line-item revenue?"

opportunity to improve on-time delivery performance. A small investment in improved availability of the top few products will significantly reduce average OCT.

In the remainder of this section, we describe the evolution of the RCO tool.

9.2.1.2 Math Programming Approaches to Optimize Coverage

The HP Labs team started by formulating the problem of finding the portfolio of size at most n that maximizes the revenue of covered orders as an integer program, IP(n):

$$\begin{aligned}
&\text{IP}(n)\text{: Maximize } \textstyle\sum_o r_o y_o \text{ subject to:}\\
&\quad (1)\ y_o \le x_p \text{ for each product-order combination } (o,p)\\
&\quad (2)\ \textstyle\sum_p x_p \le n\\
&\quad (3)\ x_p \in \{0,1\}, \quad y_o \in \{0,1\},
\end{aligned}$$

where r_o is the revenue of order o, and binary decision variables x_p and y_o represent whether product p is included in the portfolio and whether order o is covered by the portfolio, respectively.

Solving this integer program can be difficult in practice. Typical data sets have hundreds of thousands of product–order combinations, leading to hundreds of thousands of constraints of type (1). The integer program can take many hours to solve, and in some very large cases cannot be solved at all due to computer memory limitations.

However, it does have the nice property that constraints (1) are totally unimodular. This observation led to the following Lagrangian relaxation, denoted by LR(λ), in which we replace constraint (2) with a term in the objective penalizing the number of products used in the solution by a nonnegative scalar λ:

$$\begin{aligned}
&\text{LR}(\lambda)\text{: Maximize } \textstyle\sum_o r_o y_o - \lambda \sum_p x_p \text{ subject to:}\\
&\quad y_o \le x_p \text{ for each product–order combination } (o,p)\\
&\quad x_p \in [0,1],\ y_o \in [0,1].
\end{aligned}$$

The Lagrangian relaxation offers several advantages over the integer program. As mentioned previously, the remaining constraints are totally unimodular and so its optimal solution to a linear program is integer. Moreover, if a set of orders and products (O, P) is the optimal solution to LR(λ), then it will be an optimal solution to the original integer program IP($|P|$).

One very nice property of the series of solutions generated by this method is that they are nested, as is shown in the proof of the following theorem. This nested property is essential to application of the approach in business decisions, where a range of alternative portfolio choices are desired. Let $O(\lambda)$ denote the set of orders covered in the optimal solution to LR(λ), and let $P(O)$ denote the set of all products appearing in at least one order in O.

Theorem 1 *If* $\lambda_1 < \lambda 2$, *then* $O(\lambda_2) \subseteq O(\lambda_1)$.

Proof Suppose $O\lambda_2 \not\subseteq O(\lambda_1)$. Then let $O' = O(\lambda_2) \backslash O(\lambda_1) \neq \emptyset$. Then

$$\begin{aligned} 0 \geq & |O'| - \lambda_1 |P(O') \backslash P(O(\lambda_1))| \\ > & |O'| - \lambda_2 |P(O') \backslash P(O(\lambda_1))| \\ \geq & |O'| - \lambda_2 |P(O') \backslash P(O(\lambda_1))|. \end{aligned}$$

The first inequality holds by the optimality of $O(\lambda_1)$ for λ_1; if this inequality were not true, then one could increase the objective function of $\mathrm{LR}(\lambda_1)$ by adding the orders in O' to $O(\lambda_1)$. The second inequality follows from the fact that $\lambda_1 < \lambda_2$. The third inequality is true because, by the definition of O', the set of orders $O(\lambda_2) \backslash O'$ is contained in $O(\lambda_1)$ and so $P(O(\lambda_2) \backslash O') \subseteq P(O(\lambda_1))$. However, if $|O'| - \lambda_2 |P(O') \backslash P(O(\lambda_2) \backslash O')| \leq 0$, then one could improve the objective of $\mathrm{LR}(\lambda_2)$ by removing O' from $O(\lambda_2)$, which contradicts the optimality of $O(\lambda_2)$ for $\mathrm{LR}(\lambda_2)$. Thus $O(\lambda_2) \subseteq O(\lambda_1)$. □

Solving $\mathrm{LR}(\lambda)$ for a series of values of λ generates a series of solutions to $\mathrm{IP}(n)$ for several values of n. These solutions lie along the efficient frontier of revenue coverage vs. portfolio size. This series does not provide an integer solution for every possible value of n; solutions below the concave envelope of the efficient frontier are skipped. However, a wise selection of values of λ produces quite a dense curve of solutions for typical HP data sets; the number of distinct solutions is typically at least 85% of the total product count. To obtain a complete product ranking, we must break ties among products that are added between consecutive solutions to $\mathrm{LR}(\lambda)$. We employ a product's *revenue impact*, the total revenue or orders containing the product, as a tie-breaking metric. This metric proved to be the best approximation to RCO among the heuristics we tried (see Figure 9.1).

Our original implementation of RCO used a linear programming solver (CPLEX) to solve the series of problems $\mathrm{LR}(\lambda)$. However, for very large problems containing millions of order line items, each such problem can take several minutes to solve. To solve it for many values of λ in order to create a dense efficient frontier can take many hours. Large problems called for a more efficient approach to solve the series of problems $\mathrm{LR}(\lambda)$.

9.2.1.3 Relationship to Maximum Flow Problem

We learned that the problem $\mathrm{LR}(\lambda)$ for fixed λ is an example of a *selection problem* introduced independently in Balinski [4] and Rhys [25]. The former paper showed that a selection problem is equivalent to the problem of finding a minimum cut in a particular bipartite network. To see how $\mathrm{LR}(\lambda)$ can be viewed as a minimum cut problem, consider the network in Figure 9.2. Adjacent to the source node s is a set of nodes, each corresponding to one product. Adjacent to the sink node t is a set of nodes, each corresponding to one order. The capacity of the links adjacent to s

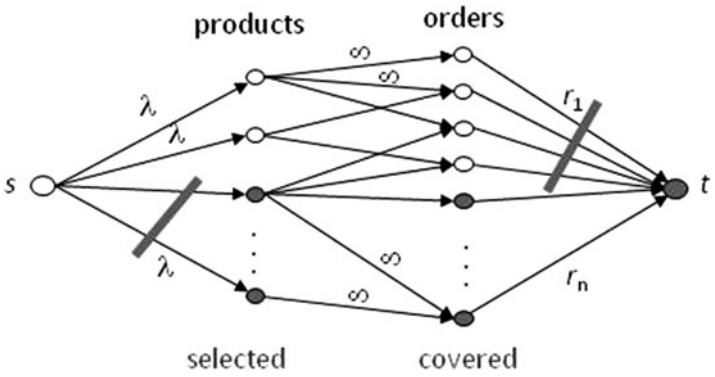


Fig. 9.2 A bipartite minimum cut/maximum flow problem corresponding to the Lagrangian relaxation LR(λ).

is λ. The capacity of the link from the node for order i is the revenue of order i. The capacity of links between product nodes and order node is infinite.

For the network shown in Figure 9.2, the set T in a minimum cut corresponds to the products selected and orders covered by an optimal solution to LR(λ). To see why, first observe that since the links from product nodes to order nodes have infinite capacity, they will not be included in a finite capacity cut. Therefore, for any order nodes in the T set of a finite capacity cut, each product that is in the order must also have its node in T. So a finite capacity cut corresponds to a feasible solution to LR(λ). Moreover, the value of an s–t cut is $\sum_o r_o(1-y_o)+\lambda\sum_p x_p$; in other words, the revenue of the orders *not* covered by the portfolio, plus λ times the number of products in the portfolio. Minimizing this quantity is equivalent to maximizing $\sum_o r_o y_o - \lambda\sum_p x_p$; therefore a minimum cut is an optimal solution to LR(λ).

It is a well-known result of Ford and Fulkerson [11] that the value of a maximal flow equals the value of a minimum cut. Moreover, the minimum cut can be obtained by finding a maximal flow.

If λ is allowed to vary, the problem LR(λ) becomes a parametric maximum flow problem, since the arc capacities depend on the parameter λ. There are several known algorithms for parametric maximum flow, such as those in Gallo et al. [12] for general networks and Ahuja et al. [1] for bipartite networks. In most prior algorithms for parametric maximum flow, a series of maximum flow problems is solved, and previous problem's solution is used to speed up the solution to the next one. By comparison, the HP Labs team developed a new parametric maximum flow algorithm for bipartite networks that finds the maximum flow for all *breakpoints* of the parameter values simultaneously (Zhang et al. [28], Tarjan et al. [30–32]). If we look at the maximum flow from the source s to the target t as a scalar function of the parameter λ, this maximum flow is a piecewise linear function of λ. A breakpoint of the parameter value is where the derivative of the piecewise linear function changes.

9.2.1.4 Parametric Bipartite Maximum Flow Algorithm

As mentioned above, the problem LR(λ) is equivalent to finding a feasible assignment of flows in the graph that maximizes the total flow from s to t. The SPMF algorithm takes advantage of the special structure of the capacity constraints.

The intuition behind the algorithm is as follows. First assume that λ=∞. Then the only constraints on flows result from the capacity limitations on arcs incident to t. It is easy to find flow assignments that saturate all capacitated links, resulting in a maximum total flow.

The next step is to find such a maximum flow assignment that distributes flows as evenly as possible across all arcs leaving s. The property "evenly as possible" means that it is impossible to rebalance flows between any pair of arcs in such a way that the absolute difference between these two flows decreases. Note that even in this most even maximum flow assignment, not all flows will be the same.

Now, with the most even assignment discussed above, impose capacity constraints of $\lambda < \infty$ on the arcs leaving s. If the flow assignment for one of these given arcs exceeds λ, reduce the flow on this arc to λ and propagate the flow reduction appropriately through the rest of the graph.

Since the original flow assignment was most evenly balanced, the total flow lost to the capacity constraint is minimal and the total flow remaining is maximal for the given parameter λ.

More formally, the algorithm works as follows:

Step 1. For a graph as in Figure 9.2 with $\lambda = \infty$, select an initial flow assignment that saturates the arcs incident to t. This is most easily done backward, starting from t and choosing an arbitrary path for a flow of size r_i from t through o_i to s.

Step 2. Rebalance the flow assignment iteratively to obtain a "most evenly balanced" flow assignment. Let $f(a \to b)$ denote the flow along the link from node a to node b. The rule for redistributing the flows is as follows. Pick i and j for which there exists an order node o_k as well as arcs $p_i \to o_k$ and $p_j \to o_k$ such that $f(s \to p_i) < f(s \to p_j)$ and $f(p_j \to o_k) > 0$. Then, reduce $f(s \to p_j)$ and $f(p_j \to o_k)$ by $\min\{(f(s \to p_j) - f(s \to p_i))/2, f(p_j \to o_k)\}$ and increase $f(s \to p_i)$ and $f(p_i \to o_k)$ by the same amount. Repeat *Step* 2 until no such rebalancing can be found.

The procedure in *Step* 2 converges, as proven in Zhang et al. [30, 31]. The limit is a flow assignment that is "most evenly balanced." In addition, since total flow is never reduced, the resulting flow assignment is a maximum flow for the graph with $\lambda = \infty$.

Step 3. To find a maximum flow assignment for a given value of λ, replace flows exceeding λ on arcs leaving the source s by λ and reduce subsequent flows appropriately to reconcile flow conservation. The resulting flow assignment is a maximum flow for λ.

For more details and a rigorous mathematical treatment of the problem, see Zhang et al. [31]. In Zhang et al. [30] it is shown that the algorithm generalizes to the case where arc capacities are a more general function of a single parameter.

In addition, since our application requires only knowledge of the minimum cut, one only needs to identify those arcs that exceed the capacity limit of λ after *Step* 2. Those arcs will be part of the minimum cut, and the ones leaving s with flows less than λ will not. To find the remaining arcs that are part of the minimum cut, one only has to identify which order nodes connect to s through one of the arcs with flows less than λ and cut through those nodes' arcs to t.

As discussed earlier, a bipartite minimum cut/maximum flow problem corresponds to the Lagrangian relaxation problem LR(λ). It can be shown that the t-partition of the minimum cut with respect to λ contains products whose flows from the source equals λ and the orders containing only those products. These products constitute the optimal portfolio for parameter λ.

Note that *Steps* 1 and 2 are independent of λ. The result of *Step* 2 allows us immediately to determine the optimal portfolio for any value of λ.

Since the flows are balanced between two arcs $s \rightarrow p_i$ and $s \rightarrow p_j$, in the algorithm described above, we call it arc-balancing method. Arc-balancing SPMF reduced the time for finding the entire efficient frontier from hours to a couple of minutes.

Another version of SPMF algorithm was developed based on the idea of redistributing the flows going into a node o in a single step so that for all pairs $p_i \rightarrow o$ and $p_j \rightarrow o$, flows $f(s \rightarrow p_j)$ and $f(p_j \rightarrow o_k)$ are "most evenly balanced." This method of redistributing flows around a vertex o is named vertex-balancing method [32]. Vertex-balancing SPMF further reduces the time for finding the entire efficient frontier to seconds.

9.2.1.5 Comparison to Other Approaches

Because the Lagrangian relaxation skips some portfolio sizes in its series of solutions, the worst-case difference between the RCO coverage and the optimal integer program's coverage can be significant. This can be illustrated through a simple example with four products and three orders shown in Table 9.1. The solutions to the integer program, Lagrangian relaxation, and RCO for this example are shown in Table 9.2. In this example, solving the Lagrangian relaxation LR(λ) for any $\lambda \in$[0,21/4] generates the portfolio {1, 2, 3, 4}; any larger value of λ yields the empty portfolio. Portfolio sizes 1, 2, and 3 are skipped and the corresponding revenue covered is zero. RCO invokes the revenue-impact heuristic to break ties among products, thereby achieving better coverage than the Lagrangian relaxation alone.

Table 9.1 A simple example of order data

Order	Products	Order Revenue
A	{1,2,3}	$12
B	{3,4}	$6
C	{1}	$3

Table 9.2 Solutions to example problem for several approaches

	Integer Program		Lagrangian Relaxation		RCO	
Portfolio Size	Solution	Revenue Covered	Solution	Revenue Covered	Solution	Revenue Covered
1	{1}	$3	skipped	$0	{3}	$0
2	{3,4}	$6	skipped	$0	{1,3}	$3
3	{1,2,3}	$12	skipped	$0	{1,2,3}	$12
4	{1,2,3,4}	$21	{1,2,3,4}	$21	{1,2,3,4}	$21

While this example illustrates worst-case behavior, in practice, RCO typically performs very close to optimal because the Lagrangian relaxation skips few solutions when applied to large order data sets from HP's business. RCO also has the added benefit of producing a nested subset of solutions, which is not true in general of the series of solutions to the integer program. Moreover, RCO compares favorably to other heuristics for ranking products (Figure 9.1). The gray curves show the cumulative revenue coverage achieved by four heuristic product rankings, in comparison to the coverage achieved by RCO. The best alternative to RCO for typical data sets is one that ranks each product according to its *revenue impact*, a metric our team devised to represent the total revenue of orders in which the product appears. The revenue-impact heuristic comes closest to RCO's coverage curve, because it is best among the heuristics at capturing product interdependencies. Still, in our empirical tests, we found that the revenue-impact ranking provides notably less revenue coverage than RCO's ranking. Given that RCO runs in less than 2 min for typical data sets and requires no more data than the heuristics, HP had no reason to settle for inferior coverage.

Another advantage of the RCO model is in its data requirements. Unlike metrics based on individual product performance, RCO does not require the metric associated with orders to be broken down to individual products in the order. This is an advantage in applying RCO to real-world data, where it is often difficult to break down an order-level metric to the product level.

9.2.1.6 Generalizations

While the discussion thus far has emphasized the application of maximizing historical revenue coverage subject to a constraint on portfolio size, this approach is flexible enough to accommodate a much wider range of objectives, such as coverage of order margin, number of orders, or any other metric associated with individual orders. It can easily accommodate up-front strategic constraints on product inclusion or exclusion. RCO can also be applied at any level of the product hierarchy, from SKUs down to components. Moreover, our algorithm has broader applications, such as in the selection of parts and tools for repair kits, terminal selection in transportation networks, and database record segmentation. Each of these problems can be naturally formulated as a parametric maximum flow problem in a bipartite network.

The SPMF algorithm has applications well beyond product portfolio management, such as in the selection of parts and tools for repair kits, terminal selection in transportation networks, and database record segmentation. The team's extension of SPMF to non-parametric max flows in general networks (Tarjan et al [28]) has an even broader range of applications in areas such as airline scheduling, open pit mining, graph partitioning in social networks, baseball elimination, staff scheduling, and homeland security.

9.2.1.7 Implementation

HP businesses typically use the previous 3 months of orders as input data to RCO, because this duration provides a representative set of orders. Significantly longer horizons might place too much weight on products that are obsolete or nearing end of life. When analysis on longer horizons is desired, RCO allows weighting of orders in the objective, thus placing more emphasis on covering the most recent orders in a given time window.

The RCO tool was not meant to replace human judgment in the design of the product portfolio. Portfolio design depends critically on knowledge of strategic new product introductions and planned obsolescence, which historical order data do not reveal. Instead, RCO is used to enhance and facilitate interactive human processes that include such strategic considerations.

9.2.2 Results

Various HP businesses have used RCO in different ways to manage their product portfolios more effectively. This section describes benefits obtained in several businesses across HP.

PSG Recommended Offering Program. PSG has used RCO to improve competitiveness by significantly reducing order cycle time. PSG used RCO to analyze order history for the USA, Europe, Middle East and Africa (EMEA), and Asia/Pacific (APJ). RCO revealed that roughly 20% of products, if optimally selected, would completely fulfill 80–85% of all customer orders. When these 20% of items are stocked to be ready-to-ship, they help significantly decrease order cycle time for a majority of orders. Using this insight, PSG established Recommended Offering for each region.

Today, the Notebook Recommended Offering ships 4 days faster than the overall Notebook offering. In EMEA, the Desktop Recommended Offering ships on average 2 days faster than the rest of the offering. The savings are impressive. Lower order cycle time improves competitiveness, each day of OCT improvement across PSG saves roughly $50M annually. PSG management estimates they have realized savings of $130M per year in EMEA and the USA. APJ is also anticipating strong benefits as they roll out the program there.

PSG Global Series Offering Program. RCO is used on an ongoing basis by the PSG Global Business Team to define the Global Series Offering for commercial notebooks. The Global Series Offering is the set of products available to HP's largest global customers. As a result of RCO, global customers are now ordering over 80% of their notebook needs from the global series portfolio, compared to 15% prior to the use of RCO. The total notebook business for global customers is $2.6B. PSG estimates the benefits of this 18% increased utilization of the recommended portfolio to be $130M in revenue.

BCS Portfolio Simplification. BCS runs RCO quarterly to evaluate its product portfolio. In the last 2 years, RCO has been used to eliminate 3,300 products from the portfolio of over 10,000 products. BCS Supply Chain Managers estimate that this reduction has resulted in $11M cost savings due to reduced inventory and planning costs. Moreover, BCS has used RCO to design options for new product platforms based on order history for previous generation platforms.

9.2.3 Summary

The coverage metric provides a new way to evaluate product portfolios. Coverage looks beyond the individual performance of products and considers their interdependence through orders, which is particularly important in configurable product businesses. This metric, and HP Labs' accompanying optimization tool, RCO, enables HP to identify products most critical to its offering, as well as candidates for discontinuance. As a result, HP has improved its operational focus on key products while also reducing the complexity of its product offering, leading to improved execution, significant cost savings, and increased customer satisfaction.

9.3 Wisdom Without the Crowd

Forecasting has been important since the dawn of business. Fundamentally, it is an exercise of using today's information to predict tomorrow's events. The popular approach, backed up by decades of development of computing technologies, is the use of statistical analysis on historical data. This approach can be very successful when the relevant information is captured in historical data.

In many situations, there is either no historical data or the data contain no useful pattern for forecasting. A good example is the forecast of the demand of a new product. A new approach is to tap into tacit and subjective information in the minds of individuals. Groups consistently perform better than individuals in forecasting future events. This so-called wisdom of crowds phenomenon has been documented over the centuries. The prediction market, where people are allowed to interact in organized markets governed by well-defined interaction rules, was shown to be an effective way to tap into the collective intelligence of crowds. Real-world examples include the Hollywood Stock Exchange and the Iowa Electronic Markets. There

are also several companies providing services of conducting prediction markets for business clients.

Prediction markets generally involve the trading of state-contingent securities. If these markets are large enough and properly designed, they can be more accurate than other techniques for extracting diffuse information, such as surveys and opinion polls. However, there are problems, particularly in the context of business forecasting. In particular, a market works when it is efficient. That is, the pool of participants is large enough, and there are plenty of trading activities. Forecasting business events, on the other hand, may involve only a handful of busy experts, and they do not constitute an efficient market.

Here, we describe an alternate method of harnessing the distributed knowledge of a small group of individuals by using a two-stage mechanism. This mechanism is designed to work on small groups, or even an individual. In the first stage, a calibration process is used to extract risk attitudes from the participants, as well as their ability to predict given outcome. In the second stage, individuals are simply asked to provide forecasts about an uncertain event, and they are rewarded according to the accuracies of their forecasts. The information gathered in the first stage is then used to de-bias and normalize the reports gathered in the second stage, which is aggregated into a single probabilistic forecast. As we show empirically, this nonlinear aggregation mechanism vastly outperforms both the imperfect market and the best of the participants. This technique has been applied to several real-world demand forecasting problems. We will present a case study of its use in demand forecasting of a technology hardware product and also discuss issues about real-world implementations.

9.3.1 Mechanism Design

We consider first an environment in which a set of N people have private information about a future event. If information across individuals is independent, and if the individuals truthfully reveal their probability beliefs, then it would be straightforward to compute the true aggregated, posterior, probabilities using Bayes' rule. If the individual i receives independent information then the probability of an outcome s, conditioned on all of their observed information I, is given by

$$P(s\,|\,I) = \frac{p_{s_1} p_{s_2} \cdots p_{s_N}}{\sum_{\text{all } s} p_{s_1} p_{s_2} \cdots p_{s_N}}, \tag{9.1}$$

where p_{s_i} is the probability that individual i predicts outcome s. This result allows us simply to take the individual predictions, multiply them together, and normalize them in order to get an aggregate probability distribution.

However, individuals do not necessarily reveal their true probabilistic beliefs. For that, we turn to *scoring rule mechanisms*. There are several proper scoring rules (for example, Brier [8]) that will solicit truthful revelation of probabilistic beliefs from risk-neutral payoff maximizing individuals. In particular we use the information entropy score. The mechanism works as follows. We ask each player to report a vector of perceived state probabilities $\{q_1, q_2, \ldots q_N\}$ with the constraint that the

vector sums to one. Then the true state x is revealed and each player paid $c_1 + c_2 \log(q_x)$, where c_1 and c_2 are positive numbers. It is straightforward to verify that if an individual believes the probability to be $\{p_1, p_2, \ldots, p_N\}$ and he or she maximizes the expected payoff, he or she will report $\{q_1 = p_1, q_2 = p_2, \ldots, q_N = p_N\}$.

Furthermore, there is ample evidence in the literature that individuals are not risk-neutral payoff maximizers. In most realistic situations, a risk-averse person will report a probability distribution that is flatter than their true beliefs as they tend to spread their bets among all possible outcomes. In the extreme case of risk aversion, an individual will report a uniform probability distribution regardless of their information. In this case, no predictive information is revealed by the report. Conversely, a risk-loving individual will tend to report a probability distribution that is more sharply peaked around a particular prediction, and in the extreme case of risk-loving behavior a subject's optimal response will be to put all the weight on the most probable state according to their observations. In this case, their report will contain some, but not all the information contained in their observations.

In order to account for both the diverse levels of risk aversion and information strengths, we add a first stage to the mechanism. Before each individual is asked to report their beliefs, their risk behavior is measured and captured by a single parameter. In the original research, and subsequent experiments that validated the effectiveness of the mechanism, we use a market mechanism, designed to elicit their risk attitudes and other relevant behavioral information. We use the portfolio held by individuals to calculate their correction factor. The formula to calculate this factor is determined empirically and has little theoretical basis.[2]

The aggregation function, after behavioral corrections, is

$$P(s \mid I) = \frac{p_{s_1}^{\beta_1} p_{s_2}^{\beta_2} \cdots p_{s_N}^{\beta_N}}{\sum_{\text{all } s} p_{s_1}^{\beta_1} p_{s_2}^{\beta_2} \cdots p_{s_N}^{\beta_N}}, \tag{9.2}$$

where β_i is the exponent assigned to individual i. The role of β_i is to help recover the true posterior probabilities from individual i's report. The value of β_i for a risk-neutral individual is one, as this individual should report the true probabilities coming out of their information. For a risk-averse individual, β_i is greater than one so as to compensate for the flatter distribution that such individuals report. The reverse, namely β_i smaller than one, applies to risk-loving individuals. The technique of soliciting this behavior adjustment parameter β_i has evolved over time. In some of the later applications, surveys were used for initial estimations and the estimates were updated using historical performance measures. Finally, a learning mechanism was used to only aggregate the best performing individuals on a moving average basis.

[2] In terms of both the market performance and the individual holdings and risk behavior, a simple functional form for β_i is given by $\beta_i = r(V_i/\sigma_i)c$, where r is a parameter that captures the risk attitude of the whole market and is reflected in the market prices of the assets, V_i is the utility of individual i, and σ_i is the variance of their holdings over time. We use c as a normalization factor so that if $r = 1$, β_i equals the number of individuals. Thus the problem lies in the actual determination of the risk attitudes both of the market as a whole and of the individual players.

9.4 Experimental Verification

A number of experiments were conducted at Hewlett-Packard Laboratories in Palo Alto, CA, to test this mechanism. Since we do not observe the underlying information in real-world situations, a large forecast error can be caused by either a failure to aggregate information or the individuals having no information. Thus, laboratory experiments, where we know the amount of information in the system, are necessary to determine how well this mechanism aggregates information. We use undergraduate and graduate students at Stanford University as subjects in a series of experiments. Five sessions were conducted with 8–13 subjects in each.

The two-stage mechanism was implemented in the laboratory setting. Possible outcomes were referred to as "states" in the experiments. There were 10 possible states, A through *J*, in all the experiments. The information available to the subjects consisted of observed sets of random draws from an urn with replacement. After privately drawing the state for the ensuing period, we filled the urn with one ball for each state, plus an additional two balls for the just-drawn true state security. Thus, it is slightly more likely to observe a ball for the true state than others. We also implemented the prediction market in the experiment, as a comparison.

The amount of information given to subjects is controlled by letting them observe different number of draws from the urn. Three types of information structures were used to ensure that the results obtained were robust. In the first treatment, each subject received three draws from the urn, with replacement. In the second treatment, half of the subjects received five draws with replacement and the other half received one. In a third treatment, half of the subjects received a random number of draws (averaging three, and also set such that the total number of draws in the community was $3N$) and the other half received three, again with replacement.

We compare the scoring rule mechanism, with behavioral correction, to three alternatives: the prediction market, reports from the best player (identified ex post, with behavioral correction), and aggregation without behavioral correction. Table 9.3 summarizes the results.

The mechanism (aggregation with behavioral correction) worked well in all the experiments. It resulted in significantly lower Kullback–Leibler measures than the no information case, the market prediction, and the best a single player could do. In fact, it performed almost three times as well as the information market. Furthermore, the nonlinear aggregation function, with behavioral correction, exhibited a smaller standard deviation than the market prediction, which indicates that the quality of its predictions, as measured by the Kullback–Leibler measure,[3] is more consistent than that of the market. In three of five cases, it also offered substantial improvements over the case without the behavioral correction.

[3] The Kullback–Leibler measure (KL measure) is a relative entropy measure, with respect to the distribution conditioned on all information available in an experiment. A KL measure of zero is a perfect match.

Table 9.3 Kullback–Leibler measure (smaller = better), by experiment

No Information	Prediction Market	Best Player	Aggregation *Without* Behavioral Correction	Aggregation *With* Behavioral Correction
1.977 (0.312)	1.222 (0.650)	0.844 (0.599)	1.105 (2.331)	0.553 (1.057)
1.501 (0.618)	1.112 (0.594)	1.128 (0.389)	0.207 (0.215)	0.214 (0.195)
1.689 (0.576)	1.053 (1.083)	0.876 (0.646)	0.489 (0.754)	0.414 (0.404)
1.635 (0.570)	1.136 (0.193)	1.074 (0.462)	0.253 (0.325)	0.413 (0.260)
1.640 (0.598)	1.371 (0.661)	1.164 (0.944)	0.478 (0.568)	0.395 (0.407)

9.5 Applications and Results

This mechanism was implemented into a web application called BRAIN (Behaviorally Robust Aggregation of Information in Networks). The process is used for forecasting tasks in several companies including a major European telecommunication company and several divisions of the largest technology company in the USA. Participants enter their reports through a web site. The behavioral corrections are carried out automatically and management can access the results directly from the web site.

A project was started in spring 2009 to make use of this process to forecast sales of a technology product. Two business events are to be forecasted. The first is the worldwide monthly shipment units of this product. This product sells into two different customer segments (designated A and B). The second is the percentage of the worldwide shipment going into customer segment A for a particular month.

For each event (for example, worldwide shipment in September 2009), there are six forecasts, two in each month for the 3 months leading up to the event. The forecasts are typically conducted in the first and third week of the month. For the September 2009 shipment, the forecasting process is conducted in late June, twice in July, twice in August and in early September. Note that partial information about shipment of September is available when the forecasting process is conducted. The design allows the forecasts to be updated if new information is available to the individuals. For each event, the real line is divided into distinct intervals and each interval is considered a possible outcome. Individuals are asked to "bet" (report) on each of the possible interval. Twenty-five individuals from different parts of the business organization, including marketing, finance, and supply chain management functions, were recruited for this process. The first forecast was conducted in late May 2009. Participation fluctuated. In the forecasts conducted in early August 2009, 16 out of the 25 recruits (64%) submitted their reports. A small budget was authorized as incentive to pay the participants.

The following figure shows the predictions and the actual events for July 2009. The predictions for Shipments and Customer Segment A have varied over the course of the predictions. The ranges are the bin widths. Prediction starts with the Early June forecasts, beginning about 7 weeks prior to the actual event.

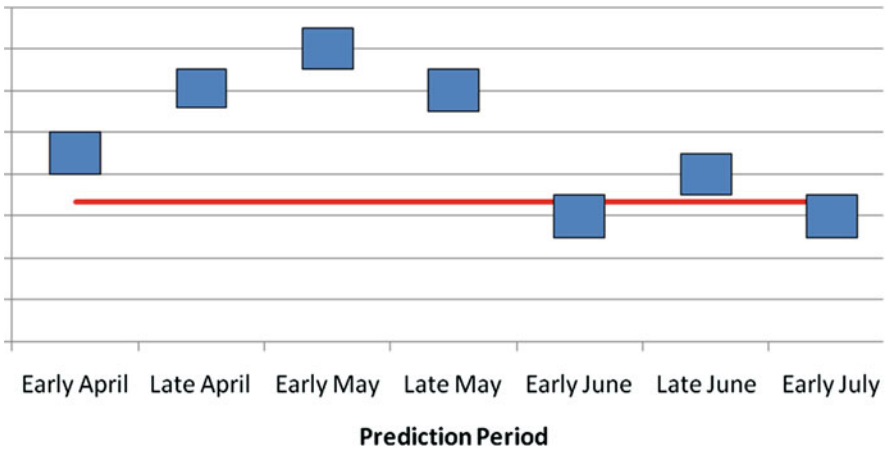


Fig. 9.3 Shipment forecast (units not available). *Note: Rectangles*: most likely interval; *thick line*: actual outcome

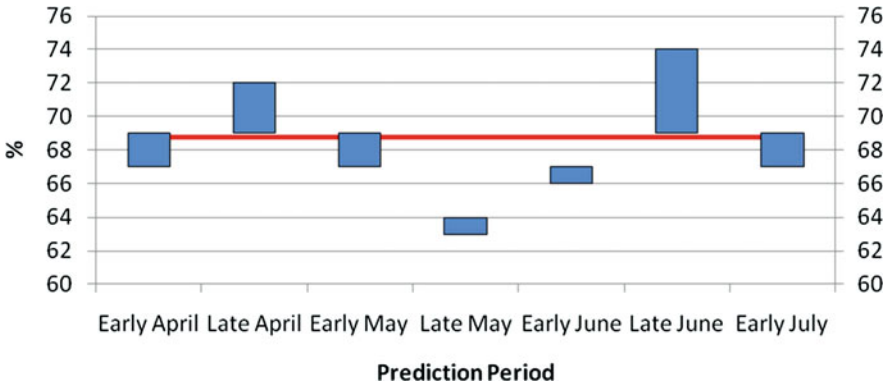


Fig. 9.4 Customer Segment A % forecast. *Note: Rectangles*: most likely interval; *thick line*: actual outcome

As one can see, the BRAIN process has provided accurate forecast at least 1 month in advance for the shipment prediction and 3 months in advance for July consumer percentage. BRAIN is also more accurate in comparison to other internal business forecasts. In particular, the shipment forecasts made 1 month prior for each month from May through July had an absolute error of 2.5% using BRAIN vs. an absolute error of 6.0% for the current forecasting method.

9.6 Modeling Rare Events in Marketing: Not a Rare Event

A rare event is an event with a very small probability of occurrence. Rare event data could be of the form where the binary dependent variable has dozens to thousands of times fewer ones ("events") than zeros ("nonevents"). Typical examples

of such events from social sciences that readily come to mind are wars, outbreak of infections, breakdown of a city's transport system, or levies. Past examples of such events from marketing are in the area of database marketing (e.g., catalogs, newspaper inserts, direct mailers sent to a large population of prospective customers) where only a small fraction (less than 1%) responded resulting in a very small probability of a response (event) [6, 18]. The examples of rare events where they occur infrequently over a period of time can be thought of as *longitudinal rare events*, while the examples where a small subset of the population responds can be thought of as *cross-sectional rare events*.

More recent examples of rare events have emerged in marketing with the advent of the Internet and digital age and the use of new types of marketing instruments. A firm can reach a large population of potential customers through its web site, display ads, e-mails, and search marketing. But only a very small proportion of those exposed to these instruments respond. For example, of the millions of visitors to a firm's web site only a handful of them click on a link or make a purchase. To make business and policy planning more effective it is important to be able to analyze and predict these events accurately.

Rare event variables have been shown to be difficult to predict and analyze. There are two sources to the problem. The first source is that standard statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events. The intuition is that there are very few values available for the independent variables to understand the circumstances that cause an event and these few values do not fully cover the tail of the logistic regression. The model infers that there are fewer circumstances under which the event will occur resulting in an underestimate. Additionally, parametric link functions such as those used for probit or logit assume specific shapes for the underlying link functions implying a given tail probability expression that remains invariant to observed data characteristics. As a result these models cannot adjust for the case when there are not enough observations to fully span the range needed for estimating these link functions. The second source of the problem is that commonly used data collection strategies are grossly inefficient for rare events data. For example, the fear of collecting data with too few events leads to data collections with huge numbers of observations but relatively few, and poorly measured, explanatory variables, such as in international conflict data with more than a quarter-million dyads, only a few of which are at war [6, 16, 18].

Researchers have tried to tackle the problem of using logistic regression (or probit) to analyze rare events data in three ways [6]. First approach is to adjust the coefficients and predictions of the estimated logistic regression model. King and Zeng [18] describe how to adjust the maximum likelihood estimates of the logistic regression parameters to calculate approximately unbiased coefficients and predictions. Second approach is to use choice-based sampling where the sample is constructed based on the value of the dependent variable. This can cause biased results (sample selection bias) and corrections must be undertaken. Manski and Lerman [21] developed the weighted exogenous maximum likelihood (WESML) estimator for dealing with the bias. Third approach is to relax the logit or probit parametric link assumptions which can be too restrictive for rare events data. Naik and Tsai

[24] developed an isotonic single-index model and developed an efficient algorithm for its estimation.

In this study we apply the second approach of choice-based sampling to discrete-choice models and decision-tree algorithms to estimate the response probabilities at the customer level to a direct mail campaign when the campaign sizes are very large (in millions) and the response rates are extremely low. We use the predicted response probabilities to rank the customers which will allow the business to run targeted campaigns, identify best and at-risk customers, reduce their cost of running the campaign, and increase response rate.

9.6.1 Methodology

9.6.1.1 Choice-Based Sampling

In a discrete-choice modeling framework sometimes one outcome can strongly outnumber the other such as when many households do not respond (e.g., to a direct mailing). Alternative sampling designs have been proposed. A case–control or choice-based sample design is one in which the sampling is stratified on the values of the response variable itself and disproportionately more observations are sampled from the smaller group. This ensures that the variation in the dependent variable is maximized with subsequent statistical analysis accounting for this sampling strategy to ensure the estimates are asymptotically unbiased and efficient [10, 21, 22].

In the biostatistical literature, case–control studies were prompted by studies in epidemiology on the effect of exposure to possible hazards such as smoking on the risks of contracting a disease condition. In a prospective study design, a sample of individuals is followed and their responses recorded. However, many disease conditions are rare and even large studies may produce few diseased individuals (cases) and little information about the hazard. In a case–control study separate samples are taken of cases and controls—individuals without the disease [27].

In the economics literature, estimation of models to understand choices for travel modes or recreation sites has used different sampling designs to collect data on consumer choices. For example, studies of participation levels and destinations for economic activities such as recreation have traditionally been analyzed using random samples of households, with either cross-section observations or panel data on repeat choices obtained from diaries. In travel demand analysis, an alternative sampling design is to conduct intercept surveys at sites. This can result in substantial reductions in survey costs and guarantee adequate sample sizes for sites of interest, but the statistical analysis must take into account the "choice-based" sample frame [23].

There is a well-developed theory for this analysis in the case of cross-section observations, where data are collected only on the intercept trip. In site choice models when subjects are intercepted at various sites, a relevant statistical analysis is

the theory of estimation from choice-based samples due to Manski and Lerman [21] and Manski and McFadden [22]. This theory was developed for situations where the behavior of a subject was observed only on the intercept choice occasion and provided convenient estimators when all sites were sampled at a positive rate. One of these estimators, called weighted exogenous sample maximum likelihood (WESML), reweights the observations so that the weighted sample choice frequencies coincide with population frequencies. A second, called conditional maximum likelihood (CML), weights the likelihood function so that the weighted sample choice probabilities average to the sample choice frequencies. The WESML setup carries out maximum pseudolikelihood estimation with a weighted log likelihood function where in conventional choice-based sampling the weights are the sampling rates for the alternatives, given by the sample frequency divided by the population frequency for each alternative. The CML setup carries out maximum conditional likelihood estimation with a log likelihood function.

However, recently sampling schemes have emerged in the literature on recreational site choice that combine interception at sites with diaries that provide panel data on intercept respondents on subsequent choice occasions. McFadden's [23] paper provides a statistical theory for these "Intercept and Follow" surveys, and indicates where analysis based on random sampling or simple choice-based sampling requires correction.

9.6.1.2 Modeling Approach

We developed a discrete-choice (logit) model and a classification-tree algorithm (aucCART) for predicting a user's probability of responding to an e-mail. The discrete-choice model is statistical based while the classification-tree algorithm is machine-learning oriented. Both response modeling methods use as input dozens of columns (or attributes) from the data sample and identify the most important (relevant) columns that are predictive of the response. By employing different types of response models for predicting the same response behavior, we were able to cross-check the models and discover predictors and attribute transformations that would be overlooked and missed in a single model. We then performed hold-out (or out-of-sample) tests on the accuracy of both methods and select the best model.

The output of each model consists of the probability that each customer will respond to a campaign and the strength of each attribute that influences this probability. We extracted about 80 explanatory attributes from the transaction and campaign databases. These may be broadly classified as (1) customer static (nontime-varying) attributes such as gender and acquisition code; (2) customer dynamic attributes just prior to the campaign, which include the recency, frequency, and monetary (RFM) attributes for customer actions, responses to previous campaigns, etc.; and (3) campaign attributes such as the campaign format and the offer type (e.g., fixed price and percentage discounts, free shipping, and freebies).

Choice-Based Sampling

A typical campaign gets very low response rate. To learn a satisfactory model, we would need thousands of responses and hence millions of rows in the training data set. Fitting models with data of this size requires a considerable amount of memory and CPU time. To solve this problem, we used choice-based sampling [21]. The idea is to include all the positive responses (Y=1) in the training data set, but only a fraction f of the non-responses (Y=0). A random sample, in contrast, would sample the same fraction from the positive responses and the negative responses. Choice-based sampling dramatically shrinks the training data set by about 20-fold when f = 0.05. To adjust for this "enriched" sample, we used case weights that are inversely proportional to f. We found that this technique yields the same results with only a very slight increase in the standard errors of the coefficients in the learned model [10].

Discrete-Choice Logit Model

The logit (or logistic regression) model is a discrete-choice model for estimating the probability of a binary response (Y=1 or 0). In our application, each user i is described by a set of static attributes $X_s(i)$(such as gender and acquisition source); each campaign j is described by a set of attributes $X_c(j)$ (such as campaign offer type and message style type); each user has dynamic attributes $X_{\mathrm{d}}(i,j)$ just before campaign j (such as recency of action, i.e., the number of days between the last action and the campaign start date). Our pooled logit model postulates

$$P\{Y(i,j)=1\} = \frac{\exp[X_s(i)\beta_s + X_c(j)\beta_c + X_d(i,j)\beta_d]}{1+\exp[X_s(i)\beta_s + X_c(j)\beta_c + X_d(i,j)\beta_d]}.$$

A numerical optimization procedure finds the coefficient vectors (β_s, β_c, β_d) that maximize the following weighted likelihood function:

$$L = \prod_{i=1}^{N} [P\{Y(i,j)=1\}]^{Y(i,j)} [1-P\{Y(i,j)=1\}]^{[1-Y(i,j)]/f},$$

where f is the choice-based sampling fraction.

Decision-Tree Learner aucCART

We developed a new decision-tree model, aucCART, for scoring customers by their probability of response. A decision tree can be thought of as a hierarchy of questions with Yes or No answers, such as "Is attribute1 $>$ 1.5?" Each case starts from the root node and is "dropped down the tree" until it reaches a terminal (or leaf) node; the answer to the question at each node determines whether that case goes to the left or right sub-tree. Each terminal node is assigned a predicted class in a way that

minimizes the misclassification cost (penalty). The task of a decision-tree model is to fit a decision tree to training data, i.e., to determine the set of suitable questions or splits.

Like traditional tree models such as CART (Classification and Regression Trees) [7], aucCART is a non-parametric, algorithmic model with built-in variable selection and cross-validation. However, traditional classification trees have some deficiencies for scoring:

They are designed to minimize the misclassification risk and typically do not perform well in scoring. This is because there is a global misclassification cost function, which makes it undesirable to split a node whose class distribution is relatively far away from that of the whole population, even though there may be sufficient information to distinguish between the high- and low-scoring cases in that node. For example, assume that the two classes, say 0 and 1, occur in equal proportions in the training data and the costs of misclassifying 0 as 1 and 1 as 0 are equal. Suppose that, while fitting the tree, one finds a node with 80% 1s (and 20% 0s) which can be split into two equally sized children nodes, one with 90% 1s and the other with 70% 1s. All these nodes have a majority of 1s and will be assigned a predicted class of 1; any reasonable decision tree will not proceed with this split since it does not improve the misclassification rate. However, when scoring is the objective, this split is potentially attractive since it separates the cases at that node into a high-scoring group (90% 1s) and a lower-scoring group (70% 1s).

A related problem is the need to specify a global misclassification cost. This is not a meaningful input when the objective is to score cases.

The aucCART method is based on CART and is designed to avoid these problems. It combines a new tree-growing method that uses a local loss function to grow deeper trees and a new tree-pruning method that uses the penalized AUC risk $R_\alpha(T) = R(T) + \alpha|T|$. Here, the AUC risk $R(T)$ is the probability that a randomly selected response scores lower than a randomly selected non-response, $|T|$ is the size of the tree, and α is the regularization parameter, which is selected by cross-validation. This method is (even) more computationally intensive than CART, in part because it runs CART repeatedly on subsets of the data and in part because minimizing the penalized AUC risk requires an exhaustive search over a very large set of sub-trees; in practice, we avoid the exhaustive search by limiting the search depth. Our numerical experiments on specific data sets have shown that aucCART performs better than CART for scoring.

9.6.2 Empirical Application and Results

9.6.2.1 Background

Customers continue to use e-mails as one of their main channels for communicating and interacting online. According to Forrester Research (2007) 94% of online customers in the USA use e-mails at least once a month. Customers also ranked opt-in

e-mails among their top five sources of advertisements they trust for product information (Forrester Research 2009). E-mail marketing has become an important part of any online marketing program. In fact, according to the 2007 Forrester Research report, 60% of marketers said that they believe marketing effectiveness of e-mail as a channel of communication will increase in the next 3 years.

An HP online service with millions of users uses e-mail marketing as one of their marketing vehicles for reaching out to its customers with new product announcements and offers. In general, each e-mail campaign is sent to all users and on a regular basis with millions of customers contacted during any specific campaign. One drawback of this "spray-and-pray" approach is the increased risk of being blacklisted by Internet Service Providers (ISPs) when they receive too many complaints. In addition to direct loss of revenue when an e-mail program is stopped early, it increases the risks of using e-mail as a regular channel for communication in the future. So the marketing team was interested in methods that would help them to identify who their best customers and "at-risk" customers were and understand what key factors are that drive customer response. This would enable them to send more targeted e-mail campaigns with relevant messages and offers.

9.6.2.2 Data Set and Variables

We selected a subset of past e-mail campaigns from the marketing campaigns database that were representative of (and similar to) the planned future campaigns. We, then, selected a subset of customers from the sent list of these past campaigns. Each campaign had a date–time and a number of attributes associated with it. The campaign date allowed us to "go back in time" and derive the user's behavioral attributes just before each of the past campaigns. We, a priori, split the customers into two customer segments based on whether they did a specific action in the past (in line with the business practice). Table 9.4 gives some descriptive statistics of the two samples.

The outcome variable, response to a campaign, indicates whether or not (1 or 0) the user responded to each of the selected campaigns. For each campaign we used the campaign database to create the campaign-specific attributes. Some examples of these attributes are the e-mail message's subject line, the format of the e-mail, the value offered in the e-mail (percentage discounts, dollar amount of free products, the

Table 9.4 Descriptive statistics of the data samples

Customer Segment	Number of Campaigns	Number of Observations (Customer campaign)	Number of Observations Choice-based Sample	Number of Customers Choice-based Sample
Action-Active	32	4.2 X	0.21 X	0.16 X
Action-Inactive	25	7.8 X	0. 39 X	0.33 X

Note: We depict the sample sizes as multiples of X to anonymize the data

type of product featured), the time-of-the-year occasion of the e-mail timing (such as Christmas shopping season).

For each customer we used the full history of their transactions since registration, available in the transaction database, to create customer-specific attributes just prior to the beginning of each campaign. These attributes included recency (how many days prior to the campaign did the user take an action), frequency (how many times in the month, quarter, or year prior to the campaign did the user undertake an action), and monetary (how much in dollars did the user spend in the month, quarter, or year prior to the campaign and in which product categories). In addition, we used data sources like the US Census Bureau and other sources of first names and gender to create a first-name-to-gender translator which predicted the probability of a person being male or female given the person's first name.

We tried all reasonable transformations of the attributes and selected the ones that yielded the best model. We determined the best transformation by investigating the residual plots for the logit model. Furthermore, the output produced from our classification tree-based aucCART algorithm (which automatically transforms some attributes) also gave us some suggestions for the most appropriate transformations. In the logit model, we selected the final set of attributes by using both forward and backward step-wise selection. In forward selection, we started with a single predictor variable (attribute) and added variables (with appropriate variable transformations) one by one, until no statistically significant variable can be added, or AIC (Akaike Information Criterion) value can be improved. In backward selection, we started with all attributes (properly transformed) included in the model and delete statistically insignificant variables one at a time, until all remaining variables are statistically significant. For the classification tree-based aucCART algorithm, variable selection was automatically performed (a built-in feature of classification tree-based algorithms).

The final data sample had several hundred thousands of data rows (each row represents a user) and approximately 80 columns (each column is an attribute describing the user at various points of time). We randomly selected 50% of the rows in the data sample as training data and the rest as testing data to evaluate the two approaches and select the best one.

9.6.2.3 Validation, Model Selection, and Results

We validated our models on holdout data sets with different customers and campaigns than the training data. Our holdout tests were designed to simulate an in-the-field application of our models to existing and new customers and new campaigns.

In addition to the two approaches outlined, various heuristics or scoring rules have been commonly used by marketing professionals to predict responses and selecting target recipients. One such heuristic for selecting recipients is by action recency, which ranks recipients by the most recent to least recent in their last action;

the more recent a user's action is, the higher the probability of responding to an e-mail the heuristic predicts. We used the action recency heuristic as the baseline of what the business is using and compare it to our two approaches.

To evaluate various rules, models, and algorithms, we needed a metric that is applicable to a wide variety of models, and that is also relevant to how the models will be used.

Figure 9.5 shows a capture curve for each model or scoring rule. The capture curve measures the percentage (Y-axis) of positive responses captured (in a holdout data set) if the model is used to select a given percentage (X-axis) of customers. The capture curves indicate that the logit model approach was the most effective in predicting and capturing customer responses to e-mails than the simple RFM method (action recency) or the decision-tree approach. For example, the logit model for action-active users is able to capture 92.1% of the campaign responses by selecting only the top 50% of the users.

The model results also indicated the strongest predictors of customer response. We are not sharing those numbers to preserve business confidentiality. In general, recency of action, the dollar amount of the user's past purchases, and the user's recorded responses to prior e-mail campaigns were significant. Additional predictors were gender, e-mail format, and offer type.

HP business group is incorporating the scoring model into their customer segmentation strategy for e-mail marketing. One of the key findings was that the business can generate 90% of the total expected response by contacting just the top 50% of users. By identifying this high-response half of its user base, they will be able to (1) tailor the message content and frequency to specific user segments based on the likelihood of response, (2) greatly increase the average response per message, and (3) reduce the total volume (and cost) of messaging.

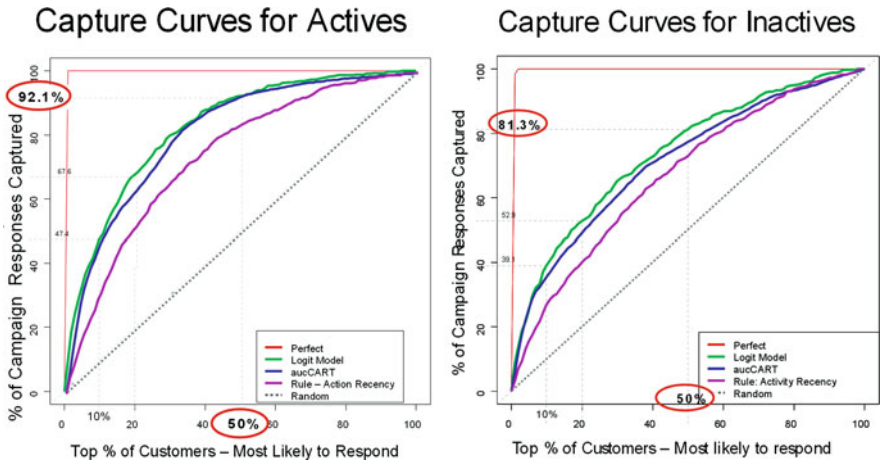


Fig. 9.5 Comparison of capture curves

In future studies we want to see if our conclusions hold for direct mail. Further we want to examine if the customers with low ranks based on the model are also the ones most likely to unsubscribe, complain, and create negative word-of-mouth.

9.7 Distribution Network Design

Hewlett-Packard provides a wide range of products and services for a diverse set of customers located across the globe leveraging a worldwide network of suppliers, partners, and facilities. As the operator of the largest supply chain in the IT industry HP relies on analytical modeling to support many strategic and operational decisions with detailed optimization models enabling evaluation of alternative supply chain strategies—procurement, location, inventory—to investigate opportunities to decrease supply chain-related costs and improve order cycle times. HP has a long tradition of employing operations research for its supply chain problems [20]. Some recent examples include reverse supply chain redesign for Personal Systems Group (PSG) in Europe [14], network design for Imaging and Printing Group (IPG) in Europe [19], production line design for IPG in the USA [9], and inventory management for former network server division [5].

In this section we describe a mathematical programming model that constitutes the core of a number of analytical decision support applications for decision problems ranging from design of manufacturing and distribution networks to evaluation of complex supplier offers in logistics procurement processes. We provide some details on two applications of the model to evaluate various distribution strategy alternatives—to answer questions such as whether it is efficient to add more distribution centers to the existing network and which distribution centers and transport modes are to be used to supply each customer location and segment—by quantifying the trade-off between the supply chain costs and order cycle times.

9.7.1 Outbound Network Design

HP provides personal computers, workstations, handheld computing devices, digital entertainment systems, calculators and other related accessories, software and services for commercial and consumer markets. The customers in the commercial segment include direct customers such as big corporations, small and medium size businesses (SMB), government agencies and indirect channel partners. Supply chain configurations vary by product as well as by customer segments. HP utilizes a number of contract manufacturers (CMs) and original design manufacturers (ODMs) to manufacture certain HP-designed products to generate cost efficiencies and reduce time to market. There are three types of nodes in a typical supply chain: inbound hubs, manufacturing sites, and outbound hubs. The inbound hubs store components and are usually situated close to the manufacturing sites. The inventory at these

locations is owned by the suppliers and is pulled by the manufacturing sites per customer order. For some critical parts, the inventory may also be owned by HP. Once the products are manufactured, they are shipped to outbound hubs (or distribution centers) for further shipment to customer locations. Certain products may also be shipped directly to customers from the manufacturing sites.

The outbound hubs play a number of critical roles in a typical supply chain. First, outbound hubs are used to consolidate shipments from manufacturing sites to customer locations for a portion of the trip. Finished goods are first shipped to an outbound hub in a bulk mode. Individual customer shipments are then scheduled for a shorter distance. Thus outbound hubs are used to leverage from volume of shipments for a particular region. Second, outbound hubs are used to merge shipments from different manufacturing sites into a single shipment to customer locations. Finally, outbound hubs are used to carry finished goods inventory for certain customers with short order cycle time requirements and for certain stable SKUs (e.g., certain standard configurations).

Given the existing and potential outbound hubs, the model is used to seek answers to the following questions: Which of the existing and potential outbound hubs should HP use in its operations? Which customer locations and segments should be assigned to each outbound hub? Which product groups should be assigned to each outbound hub? What should be the mode of transportation in meeting customer demands for each customer location and segment?

The answers to these questions hinge on various aspects of the fundamental tradeoffs between customer service levels and supply chain-related costs. The former is measured by Order to Delivery Time (ODT), the time between customer order and order delivery to customer. The latter, supply chain costs, fall in four broad categories: First, Inventory-Driven Costs (IDC) include all of the costs that derive from the level of inventory in the regions, such as obsolescence and component devaluations. Second, Trading Expenses (TE) include freight, duties, taxes, allocations, and warehousing. Third, Manufacturing Expenses (MOH) include the cost of manufacturing products, as well as any costs related to the support of that manufacturing activity including customization and rework. Finally, Cash to Cash (C2C) takes into account how long inventory is held in the region and how long it takes to pay the suppliers and to receive payment from customers.

ODT is an important metric for a product division's supply chain. Service level agreements with customers usually involve explicit ODT requirements. ODT is composed of several components such as order entry time, material wait time, factory cycle time, and delivery time. Of these components, material wait time and the delivery time are likely to get impacted by the supply chain configuration. Furthermore, for a given supply chain configuration, the three components of ODT—order entry, factory cycle time, and the delivery time—are not likely to change from one order to another (for the same customer location and product group), while the material wait time can be considerably variable depending on the immediate availability of the components at the designated inbound hub. Also note that from the above four components, delivery time is the only component that will be impacted by the outbound strategy. Different customer groups—corporate, small and medium

businesses, public sector, indirect channel partners—may have distinct ODT requirements. Any outbound strategy should ensure that the ODT requirements are satisfied for each customer segment.

Trading expenses and inventory-driven costs are likely to be impacted most by the outbound strategy. Major components of Trading Expenses are transportation costs from manufacturing sites into the outbound hubs and from outbound hubs to the customer locations, material handling costs, and facility costs. Main elements of Inventory-Driven Costs are costs due to inventory in transit from manufacturing sites to the outbound hubs, and from outbound hubs to customer locations, and inventory in the outbound hubs.

The decision problem is to minimize trading expenses and inventory-driven costs while satisfying order to delivery time targets set by management. Various business constraints such as limiting the total number of outbound hubs that will be used, forcing a particular outbound hub to stay open or closed will also need to be incorporated as constraints in the model.

Products can be modeled at the SKU level or at the product category level after aggregation. Customer segments are modeled separately as shipment volumes and ODT requirements vary by segment. For customer locations various levels of aggregation—by state, zip code, etc.—are possible. HP works with many different transportation service providers including parcel carriers, airfreight companies, less-than-truckload (LTL) and full truckload (FTL) carriers. Transportation mode can be modeled using the physical mode of transportation (type of vehicle or type of company) or using delivery times to code transportation modes—e.g., 1-day service, 2-day service, 3-day service.

In order to capture the variability in ODT targets, two modes of delivery are defined. For a fraction θ^r_{js} of orders originating from customer segment s for product j, the order needs to be shipped from the factory with regular delivery within w^r_{js}. Likewise, for a fraction $\theta^e_{js} = 1 - \theta^r_{js}$ of orders originating from customer segment s for product j, the order needs to be shipped from the factory with emergency delivery within w^e_{js}.

9.7.2 A Formal Model

We next introduce the notation needed for a formal presentation of the mathematical model. Let M denote the set of manufacturing sites, I denote the set of potential outbound hub sites, K denote the set of customer locations, S denote the set of customer segments, J denote the set of product groups, and T denote the set of transportation modes available.

The following variables define the parameters of the model:

d_{ksj} : demand in location k for customer segment s for product j

c_{mitksj} : cost to satisfy demand in location k for customer segment s for product j by manufacturing in site m through outbound hub i with transport mode t

ℓ_{mitksj} : delivery time to satisfy demand in location k for customer segment s for product j by manufacturing in site m through outbound hub i with transport mode t
f_i : fixed operating cost of outbound hub site i
C_i : capacity of outbound hub site i
w^r_{js} : time window specified for product j for customer segment s for regular delivery
w^e_{js} : time window specified for product j for customer segment s for emergency delivery
θ^r_{js} : fraction of orders in segment s for product j requiring regular delivery
θ^e_{js} : fraction of orders in segment s for product j requiring emergency delivery

We also define the following variables to be used in the mathematical program:

$$\delta^r_{mitksj} = \begin{cases} 1 & \text{if } \ell_{mitksj} \le w^r_{sj} \\ 0 & \text{otherwise} \end{cases}$$

$$\delta^e_{mitksj} = \begin{cases} 1 & \text{if } \ell_{mitksj} \le w^e_{sj} \\ 0 & \text{otherwise} \end{cases}$$

The following parameters are used to enforce a specific scenario for the network design:

$$\alpha_i = \begin{cases} 1 & \text{if outbound hub } i \text{ needs to be open in a scenario} \\ 0 & \text{otherwise} \end{cases}$$

$$\beta_i = \begin{cases} 1 & \text{if outbound hub } i \text{ needs to be closed in a scenario} \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma_i = \begin{cases} 1 & \text{if outbound hub } i\text{'s capacity needs to be enforced in a scenario} \\ 0 & \text{otherwise} \end{cases}$$

The following variables are the decision variables of the problem:

$$x^r_{mitksj} = \begin{cases} 1 & \text{if segment } s\text{'s regular demand in location } k \text{ for product } j \text{ is satisfied by manufacturing site } m \text{ through outbound hub } i \text{ with mode } t \\ 0 & \text{otherwise} \end{cases}$$

$$x^e_{mitksj} = \begin{cases} 1 & \text{if segment } s\text{'s emergency demand in location } k \text{ for product } j \text{ is satisfied by manufacturing site } m \text{ through outbound hub } i \text{ with mode } t \\ 0 & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if outbound site } i \text{ is used} \\ 0 & \text{otherwise} \end{cases}$$

With the notation introduced the mathematical program is written as

$$\min \sum_{m\in M}\sum_{i\in I}\sum_{t\in T}\sum_{k\in K}\sum_{s\in S}\sum_{j\in J} d_{ksj}c_{mitksj}[\theta^r_{sj}x^r_{mitksj}+\theta^e_{sj}x^e_{mitksj}] + \sum_{i\in I} f_i y_i, \tag{9.3}$$

$$\sum_{m\in M}\sum_{i\in I}\sum_{t\in T}\delta^r_{mitksj}x^r_{mitksj} = 1 \text{ for all } k\in K, s\in S, j\in J, \tag{9.4}$$

$$\sum_{m\in M}\sum_{i\in I}\sum_{t\in T}\delta^e_{mitksj}x^e_{mitksj} = 1 \text{ for all } k\in K, s\in S, j\in J, \tag{9.5}$$

$$x^r_{mitksj} - y_i \le 0 \text{ for all } m\in M, i\in I, t\in T, k\in K, s\in S, j\in J, \tag{9.6}$$

$$x^e_{mitksj} - y_i \le 0 \text{ for all } m\in M, i\in I, t\in T, k\in K, s\in S, j\in J, \tag{9.7}$$

$$y_i \ge \alpha_i \text{ for all } i\in I, \tag{9.8}$$

$$y_i \le (1-\beta_i) \text{ for all } i\in I, \tag{9.9}$$

$$\gamma_i\left(\sum_{m\in M}\sum_{t\in T}\sum_{k\in K}\sum_{s\in S}\sum_{j\in J} d_{ksj}[\theta^r_{sj}x^r_{mitksj} + \theta^e_{sj}x^e_{mitksj}]\right) \le C_i \text{ for all } i\in I, \tag{9.10}$$

$$x^r_{mitksj} \in \{0,1\} \text{ for all } m\in M, i\in I, t\in T, k\in K, s\in S, j\in J, \tag{9.11}$$

$$x^e_{mitksj} \in \{0,1\} \text{ for all } m\in M, i\in I, t\in T, k\in K, s\in S, j\in J, \tag{9.12}$$

$$y_i \in \{0,1\} \text{ for all } i\in I. \tag{9.13}$$

The objective in (9.3) minimizes all incoming and outgoing transportation costs, material handling, inventory, and the facility costs. The constraints in (9.4) and (9.5) ensure that each customer segment in each location is assigned to one outbound site, manufacturing site, and one transportation mode for each product group that are within delivery time requirements for regular and emergency demands, respectively. Note that the product groups from a single customer location and segment can be assigned to different manufacturing sites, outbound hubs, and transportation modes. The constraints in (9.6) and (9.7) ensure that service from an outbound hub is available only if the facility is open. The constraints in (9.8) and (9.9) ensure that the outbound hubs are forced to be open or closed based on the scenario specification.

The constraints in (9.10) ensure that the capacity of the outbound hub is enforced if specified in the scenario. The constraints in (9.11), (9.12), and (9.13) ensure that all decision variables are binary. Note that the formulation in (9.3)–(9.13) assumes that the model is full, e.g., every customer location has demand from all $|S|$ segments and for all $|J|$ product groups. This is only to make the exposition simple. The actual model used in implementation takes advantage of the link sparsity.

9.7.3 Implementation

The model in the previous section was implemented using ILOG's OPL Studio and solved using CPLEX. The raw data are stored in several tables in a database and can be imported from a spreadsheet or a flat file. A typical implementation may involve up to 1,000 customer locations (actual customer locations aggregated at the 3 digit zip code), 5–10 customer segments, 5–10 transport modes, up to 10 product groups, and up to 100 potential outbound hub sites. Standardized forms are used to allow the user to specify the parameters for what–if scenarios in order to see the impact

of several critical variables. Through various forms the user can change the delivery time targets, enforce a particular outbound hub to stay open or closed, activate or deactivate the capacity constraint on a particular hub, and limit the number of outbound hubs. The user sees the results of the model via several reports. Location Summary report shows which outbound hubs are open and what costs are incurred in doing so. Location Usage report shows the total number of units in each product category that flows through each outbound hub. Delivery Performance report shows the resulting average delivery times for each customer segment and product group. Location Customer Assignment report shows the detailed assignment of customer locations/customer segments to manufacturing sites/outbound hubs/transportation modes.

9.7.4 Regarding Data

The data requirements can be categorized into four groups: logistics, financial, demand, and customer service requirements. Some critical data elements need to be estimated from various data sources. Transportation costs and times between manufacturing sites and outbound hubs are estimated assuming that a bulk mode is used and scale economies are fully utilized, considering the typically large volume of shipments. Based on the manufacturing scenario, the shipments can be originating from various locations worldwide. Depending on the origin, the shipments may be made over the ocean by major carriers or by FTL carriers. Cost and time estimates are created using data on rate tables and maps from major carriers. Estimation of transportation costs and times between outbound hubs and customer locations is based on data on shipment histories and representative carrier cost and time information for various weight categories. Annual demands at the product group, customer segment, and customer locations were estimated from data on shipment history. In addition to these three items, data such as material handling and facility setup costs for outbound hubs, unit manufacturing costs (estimates) at different manufacturing sites, inventory holding cost rates and customer service level requirements are obtained from various sources in finance, logistics, and procurement operations.

9.7.5 Exemplary Analyses

The outbound model proved to be very useful for internal consulting teams for evaluating alternative distribution strategies for various product groups. The outbound model was also used as a primary input to the assessment of end-to-end manufacturing scenarios for a product group.

The model described above provides the core for analysis of a number of broader supply chain strategy decisions including the selection for manufacturing sites. The analysis for the outbound strategy clearly depends on the locations of the

manufacturing facilities. For this purpose, viable scenarios included the baseline scenario describing the manufacturing locations at the time of implementation. These manufacturing scenarios specify manufacturing location(s) for each product category.

For each manufacturing scenario, various analyses can be carried out. The first category of analysis takes the current level of OTD targets as input and develops an outbound strategy for each manufacturing scenario. The analysis in this category was used to determine the optimal outbound hub locations and to assess the value of additional outbound hubs for each manufacturing scenario. The analysis proved very useful in understanding the marginal value of each additional outbound hub. An example of this analysis (with fictitious data) is provided in Figure 9.6.

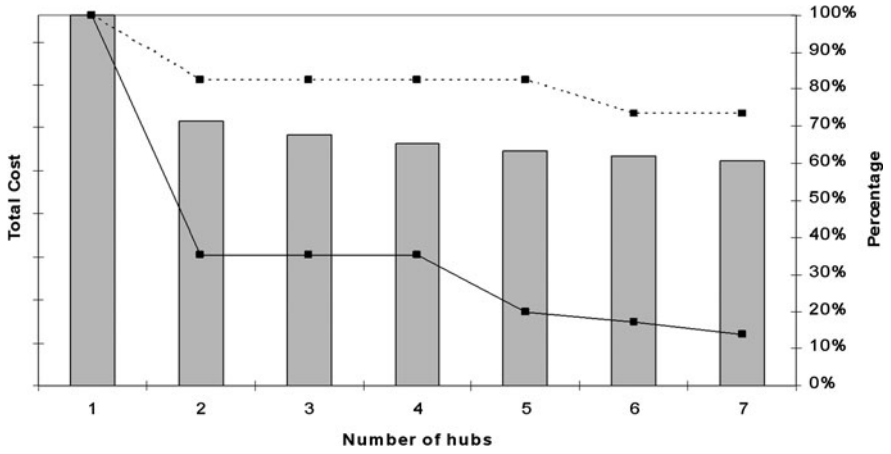


Fig. 9.6 Impact of number of hubs

In the example, we consider a manufacturing scenario with a single manufacturing location co-located with one of the outbound hubs. Since each additional hub provides the flexibility to consolidate a portion of the trip for shipments to customer locations, the total costs decline. However, as expected, there are decreasing marginal returns. Understanding the exact value of each additional outbound hub, together with an evaluation of operational complexity, provides valuable guidance for the management decisions on the number and locations of each outbound hub. In Figure 9.6, we also show the percentage of products shipped directly from the outbound hub co-located with the single manufacturing site for two product groups: bulky products (product group 1) and small/light products (product group 2). Clearly, shipment consolidation is more beneficial for bulkier items (product group 1), and we see that more of this group of products are shipped via additional hubs than the second group. The analysis is also useful in estimating the transportation cost component of different manufacturing scenarios. In addition to the strategic insights, the analysis can also be used to support detailed operational decisions such as which manufacturing sites, outbound hubs, and

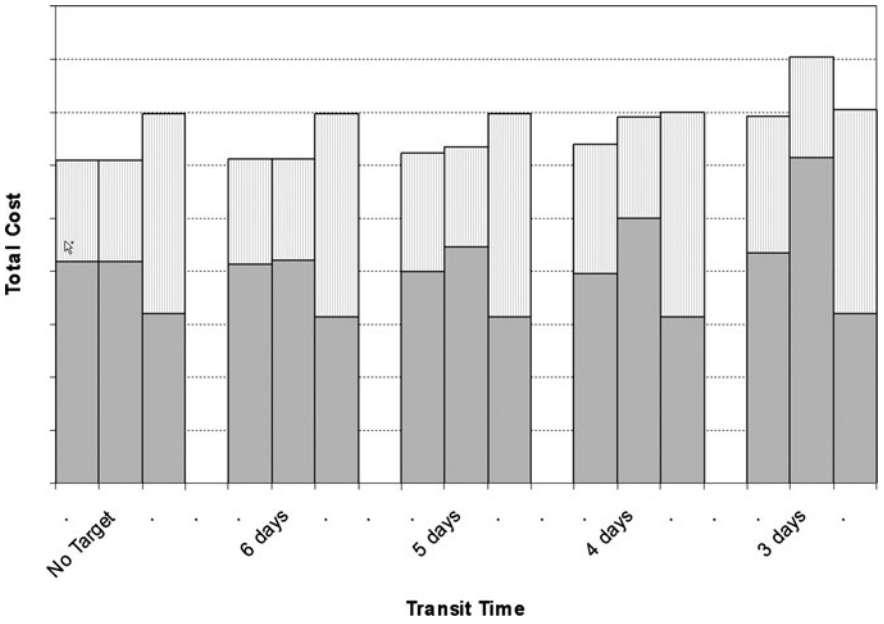


Fig. 9.7 Impact of delivery time targets

transportation modes should be used to deliver orders of a particular region and a customer segment.

The model can also be used to study the trade-off between the supply chain costs and OTD targets. Naturally, this trade-off varies with manufacturing scenarios. In Figure 9.7 we present an example of such analysis (with fictitious, but representative data). In this example, we consider three manufacturing scenarios. In the first scenario, the products can be manufactured in an offshore location as well as locally in the USA. Modifying the per unit costs c_{mitksj} in the mathematical model to include manufacturing costs, the model is used to select a manufacturing site among the set of possible sites for satisfying demand in a particular customer location. In the second scenario, all manufacturing is done at an overseas site. Finally in the third scenario, all manufacturing is done locally in the USA. Each manufacturing scenario is combined with various target delivery time levels starting from a case where there is no time constraint on shipping a customer order.

The analysis reveals that, for all three manufacturing scenarios, the total costs increase as the delivery time targets are more aggressive. The mixed scenario outperforms the other two scenarios for all target levels, since it has the flexibility of employing offshore as well as local manufacturing. For this scenario, as the delivery time targets get more aggressive, more manufacturing is moved to domestic sites. Note also that while the total cost for the offshore-only scenario is quite sensitive to the delivery time targets, the total cost for the US-only scenario is rather insensitive.

9.8 Collaborations and Conclusion

For development and deployment of decision sciences solutions, the HP Labs team works very closely with business units. In addition, in most cases HP's Information Technology (HP IT) group has a significant role in the success of the projects. The HP Labs team takes the ownership of development of underlying algorithms and core algorithmic software engine. HP IT is generally responsible for integration of the core analytical engine with back-end IT systems, database design and development, system architecture, deployment, and support of the complete system.

Over the years, HP Labs' Business Optimization Lab has built strong research collaborations with leading faculty members in several areas of interests to HP and the academic community. The university collaboration for the work presented in this chapter is reflected in the author list.

This chapter covers a very narrow slice of advanced analytics project at HP Labs and at HP. It is safe to say that the creation and application of rigorous mathematical models is well established throughout the company. Applied researchers and practitioners are making contributions that directly impact the top and bottom line.

Acknowledgments

In this chapter, we have summarized the work of several members of the HP Labs and business units of HP. In particular, we are very thankful to Kemal Guler for organizing the content of distribution network design portion of this chapter.

References

1. Ahuja RK, Orlin RB, Stein C, Tarjan RE (1994) Improved algorithms for bipartite network flow. SIAM Journal of Computing 23:903–933
2. Ansari A, Mela CF (2003) E-customization. Journal of Marketing Research XL:131–145
3. Babenko M, Derryberry J, Goldberg A, Tarjan R, Zhou Y (2007) Experimental evaluation of parametric max-flow algorithms. Proceedings of WEA. Lecture Notes in Computer Science 4525. Springer, Berlin–Heidelberg, Germany, pp. 612–623
4. Balinski ML (1970) On a selection problem. Management Science 17(3):230–231
5. Beyer D, Ward J (2002) Network server supply chain at HP: A case study. In: Song J, Yao D (eds) Supply chain structures: Coordination, information and optimization. International Series in Operations Research and Management Science, Kluwer, Norwell, MA
6. Blattberg RC, Kim P, Neslin S (2008) Database marketing: Analyzing and managing customers. Springer, New York
7. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall, New York
8. Brier, GW (1950) Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78:1–3
9. Burman M, Gershwin SB, Suyematsu C (1998) Hewlett-Packard uses operations research to improve the design of a printer production line. Interfaces 28:24–36

10. Donkers B, Franses PH, Verhoef PC (2003) Selective sampling for binary choice models. Journal of Marketing Research XL:492–497
11. Ford LR, Fulkerson DR (1956) Maximum flow through a network. Canadian Journal of Mathematics 8:339–404
12. Gallo G, Grigoriadis MD, Tarjan RE (1989) A fast parametric maximum flow algorithm and applications. SIAM Journal of Computing 18:30–55
13. Goldberg AV, Tarjan RE (1986) A new approach to the maximum flow problem. Proceedings of the 18th Annual ACM Sympos Theory Computation (Berkeley, CA), May 28–30, pp. 136–146
14. Guide Jr VDR, Mulydermans L, Van Wassenhove LN (2005) Hewlett-Packard company unlocks the value potential from time-sensitive returns. Interfaces 35:281–293
15. Jain S (2008) Decision sciences—A story of excellence at Hewlett-Packard. OR/MS Today, April
16. Kamakura WA, Mela CF, Ansari A, Bodapati A, Fader P, Iyengar R, Naik P, Neslin S, Sun B, Verhoef P, Wedel M, Wilcox R (2005) Choice models and customer relationship management. Marketing Letters 16(3/4):279–291
17. Kamakura WA, Russell GJ (1989) A probabilistic choice model for market segmentation and elasticity structure. Journal of Marketing Research 26(4):379–390
18. King G, Zeng L (2001) Logistic regression in rare events data. Political Analysis 9(2):137–163
19. Laval C, Feyhl M, Kakouros S (2005) Hewlett-Packard combined or and expert knowledge to design its supply chains. Interfaces 35:238–247
20. Lee HL, Billington C (1995) The evolution of supply chain management models and practice at Hewlett-Packard company. Interfaces 25:42–46
21. Manski CF, Lerman SR (1977) The estimation of choice probabilities from choice based samples. Econometrica 45(8)(November):1977–1988
22. Manski CF, McFadden D (1981) Structural analysis of discrete data with econometric applications. MIT, Cambridge, MA
23. McFadden D (1996) On the analysis of "Intercept and Follow" surveys. Working Paper. University of California, Berkeley, CA
24. Naik PA, Tsai CL (2004) Isotonic single-index model for high-dimensional database marketing. Computational Statistics & Data Analysis 47(4):775–790
25. Rhys JMW (1970) A selection problem of shared fixed costs and network flows. Management Science 17(3):200–207
26. Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. Marketing Science 15(4):321–340.
27. Scott AJ, Wild CJ (1986) Fitting logistic models under case-control or choice based sampling. Journal of the Royal Statistical Society 48(2):170–182
28. Tarjan R, Ward J, Zhang B, Zhou Y, Mao J (2006) Balancing applied to maximum network flow problems. Proceedings of the ESA, Lecture Notes in Computer Science 4168, pp. 612–623
29. Ward J, Zhang B, Jain S, Fry C, Olavson T, Mishal H, Amaral J, Beyer D, Brecht A, Cargille B, Chadinha R, Chou K, DeNyse G, Feng Q, Padovani C, Raj S, Sunderbruch K, Tarjan R, Venkatraman K, Woods J, Zhou J (2010) HP transforms product portfolio management with operations research. Interfaces 40(1):17–32
30. Zhang B, Ward J, Feng Q (2004) A simultaneous parametric maximum flow algorithm for finding the complete chain of solutions. HP Technical Report: HPL-2004-189, Palo Alto, CA
31. Zhang B, Ward J, Feng Q (2005a) Simultaneous parametric maximum flow algorithm for the selection model. HP Technical Report HPL-2005-91, Palo Alto, CA
32. Zhang B, Ward J, Feng Q (2005b) Simultaneous parametric maximum flow algorithm with vertex balancing, HP Technical Report HPL-2005-121, Palo Alto, CA

Chapter 10
Global Trade Process and Supply Chain Management

Hau L. Lee

Abstract As a result of increased globalization of industrial supply chains, effective supply chain management requires sound alignment with the global trade processes. The design of the global supply chain and the determination of the right level of postponement are both tied intimately to the prevailing network of trade agreements, regulations, and local requirements of the countries in which the company is operating in. Moreover, the dynamic changes and uncertainties of these agreements and requirements must be anticipated. In addition, the complexity of the cross-border trade processes results in uncertainties in the lead time and costs involved in global trade, which naturally forms part of the consideration of global sourcing, and the resulting safety stocks or other hedging decisions. Governments, exporters, importers, carriers, and other service providers have to work together to reduce the logistics frictions involved in the global trade processes. The benefits accrue not only to the exporters, importers, and the intermediaries but ultimately they could foster bilateral trade. The only way to reduce the frictions is to gain a deep understanding of the detailed process steps involved to improve upon it by using information technologies and potentially re-engineer the processes. But the payoffs to such investments can be huge. This chapter provides some preliminary discussion of the inter-relationships between global trade processes and supply chain management, with the objective to stimulate research in this area.

Prelude

Supply chain management has been my research focus for most of my professional career. I started research in this area in the last year of my PhD program at Wharton. My advisor, Professor Morris Cohen, and I were looking at how companies should structure their supply chain network, i.e., where to locate their manufacturing and distribution centers and how customers were to be served by this network. We later extended considerations of this problem when the network was global. The very first

Hau L. Lee
Graduate School of Business, Stanford University, Stanford, CA 94305, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_10,

step of our research involved a thorough literature review, and at that time, the most important and relevant paper that impacted the way we thought of our research approach was the paper by Geoffrion and Graves [5], as well as some follow-on papers by the same authors. In a way, Geoffrion and Graves [5] was the starting point of my supply chain research. It was a great pleasure to me that I eventually came to know Professor Geoffrion, the person behind the very paper that was the anchor point of my early supply chain research. I also learnt about the many contributions that he has made to the OR/MS field. It was therefore a great honor for me to contribute to this book to show our respect, admiration, and recognition of Professor Art Geoffrion. It is also very fitting that I use this opportunity to write about supply chain design and global trade, a topic that is quite linked to the Geoffrion and Graves [5] article.

10.1 Introduction

For most industries, supply chains today are increasingly global, with suppliers, manufacturers, distributers, and retail markets located globally. Companies are looking for new and perhaps lower cost sources of supply and production, new partners to innovate and develop new products, and expanding the markets to new and emerging economies. Outsourcing and offshoring have become key focal points for management. Indeed, as a recent study by Accenture showed that companies were increasingly sourcing from and selling to markets outside of where the core of the companies resided (see Figure 10.1). The globalization of supply chains consequently led to an explosion of world trade, since raw materials, components, semi-finished products, and finished products flowing through the global supply chain would need to cross country borders many times. Indeed, in the last 10 years, the

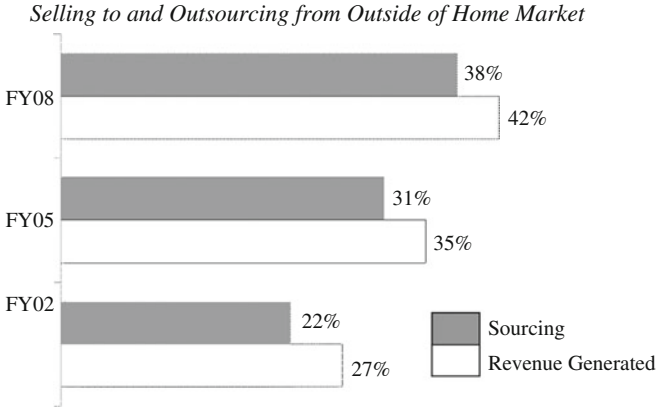


Fig. 10.1 Global sourcing and selling

volume of global trade has increased by almost 6% annually, while the growth of global GDP has been only 3% annually.

While global trade has exploded over the years, barriers to trade and other protectionism measures have also skyrocketed. Figure 10.2 shows the exponential growth of regional trade agreements globally. The increase of such trade agreements means that many countries have set up special rules and regulations for some specific products with some specific trading partners. Although these trade agreements often mean lower customs or special treatment if some requirements are met for some specific products and trading partners, their existence also means that higher customs and more restrictions are then created for other products and other trading partners.

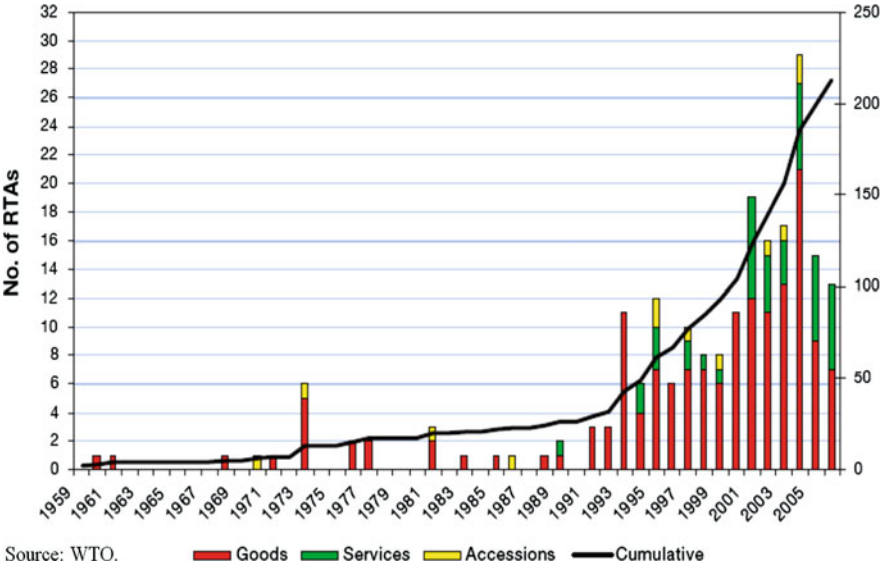


Fig. 10.2 Increasing regional trade agreements

There are several implications for both researchers and practitioners as a result of the increasing globalization of supply chains. I will focus on two. First, the design of the supply chain is a complex decision. Take sourcing as an example. Companies often compute the "total landed cost" of the various alternative sources for evaluation. The total landed cost consists of the cost of acquisition, freight cost, customs and duties, transaction costs, other logistics costs (such as documentation), potential tax subsidies, and inventory holding costs. But customs and duties is a very complex factor, as it depends on the trade agreements between the exporting and importing countries, the trade policies of these countries, duty drawback allowances, and, as we will show later, potentially all the trade agreements among all the countries involved in the supply chain of the product. We have already seen the escalation of regional trade agreements globally, making customs consideration a very difficult

and complex one. The additional challenge is that there are still a lot of uncertainties involved—the trade agreements, regulations, and requirements may change over time.

Moreover, given the trade agreements and the customs duties in place, companies should design the right postponement strategies so that the right level of customization of the product can occur in the market regions instead of centrally at the factory. The key question often is what should be the portion of the product to be built at the factory and what should be the portion to be built at the (multiple and distributed) distribution centers that are in the market regions. There are certainly economies of scale and easier production and quality control to build products at the factory; but the increasing protectionism and resulted tariffs for products can motivate having some portion of the product, i.e., some of the customization steps to be carried out in distribution. The right level of postponement must take account of the associated customs and duties implications.

Second, cross-border trade processes are non-trivial. From the initiation of export/import, to the physical movement of goods across the borders, and then the arrival at the final destination, the trade process can be complex, time-consuming, and costly. This process can also become even more complex when some nations, such as the United States, are concerned with the threat of security when container shipments can be used by terrorists as a weapon of mass destruction. The result is added documentation requirements, inspection, and delays. In the total landed cost analysis, the logistics and transaction costs, and the inventory holding costs, can be greatly affected by the cross-border trade processes. For example, if such processes are long and unreliable, then the inventory in transit will be high, and the safety stocks that the importing company need to carry would have to be increased.

In this chapter, I do not attempt to give a complete treatment of these two implications. Instead, I like to highlight some important considerations that practitioners should pay attention to, and some possible research that can be carried out.

10.2 Supply Chain Design and Trade Processes

10.2.1 Supply Chain Design

In classic supply chain design problems, the key factors for considerations include fixed and variable costs of the sites for manufacturing and distribution, the transportation costs, and inventory costs (see the classic work of Geoffrion and Graves [5]). There is a rich literature on this topic, ranging from deterministic applications (e.g., [1]) and stochastic demand versions (e.g., [2]). Extending this to global supply chain design often requires modeling additional factors such as local content requirements, customs and duties rates, differential tax rates in different countries, transfer pricing schemes, and in some cases, exchange rate fluctuations (e.g., [3]).

The proliferation of trade agreements has added a new dimension of complexity to the supply chain design problem. Prior to the Agreement on Trade-Related

Investment Measures (TRIMS Agreement), in an effort to increase labor participation in their country and attract investment, some developing countries had included particular rules that provide incentives for companies in a particular industry to enter the respective countries. These incentives often included duty-free rates or a reduction in duties paid on imports; the result was that companies increased their use of local contents in the final exported product or so-called local contents and trade balancing requirements. Under these regimes, companies could only achieve reduced duties on their imports used to serve the domestic market, by increasing the country's exports. Otherwise, companies were forced to use local contents to serve the domestic market and, in some cases, it was not possible and too costly to source the parts that the company needed. The use of the previously mentioned incentive schemas was prohibited when the TRIMS Agreement of the Multilateral Agreements on Trade in Goods was negotiated during the Uruguay Round of WTO negotiations and came into force in 1995. This agreement applied to trade in goods and generally prohibited trade-related investment measures.

Such agreements can complicate the supply chain design problem. At the same time, clever exploitation of such agreements can lead to significant savings to the firm! It was due to such duty savings that Crocs used to manufacture its plastic shoes in Canada to serve the demands in Israel, since the special trade agreement between Canada and Israel resulted in zero duty for shoes made in Canada, versus 40% otherwise [8]. Consider the Logan car of Renault (see [12]). The Logan was designed as a car for new markets with high potential growth. Renault initially targeted customers in Colombia, Iran, Romania, Russia, and the Maghreb region. However, the Logan was also designed to be sold in markets throughout Africa, Asia, Eastern Europe, and South America.

Automobiles sold in a given country could be built with a range of local contents. At one extreme, a company could export a car to its customers abroad as a completely built-up vehicle (CBU), where the importing country received a fully assembled vehicle ready for sale in the local market. CBU were advantageous in that all vehicle production and assembly could be centralized. Only the logistics would thus be required to transport the vehicle from its origin to its destination. However, duties on vehicles imported as CBUs were traditionally exorbitant, ranging from 35% in South America to 90% in Iran and 100% in India. An alternative approach was to export vehicles as completely knocked down units (CKDs). While the definition could vary by importing country where the final assembly took place, CKDs described the entire kit of parts that would be required to assemble the final vehicle.

Consider the Pitesti plant in Romania as a production site of the Logan. The Pitesti factory could be used to support Logan assembly plants in Russia, Morocco, Colombia, Brazil, India, and Iran by providing them with CKD parts. Romania could also produce Logan as CBUs for export to European countries, Croatia, and Turkey, where customs unions or free trade agreements allowed for duty-free import of CBUs, by following the rules of origin. When Romania joined the European Union (EU) in 2007, a whole new set of trade agreements became effective, and the customs and duties implications to the supply chain turned out to be huge.

The immediate consequence of Romania entering the EU was that Renault could now import parts from several countries using the free trade agreements available as a member of the EU. This was also the case for mechanical parts supplied from Brazil, which were subject to an MFN duty rate of 30% prior to 2007. Accordingly, Romanian vehicles would be considered as European vehicles. For example, the duty rate on the import of vehicles into Mexico was 50% (before Romania's accession) and 0%, with a certificate of origin (after accession).

Another outcome of Romania's membership in the EU was that, under rules of origin requirements for CBU imports into the EU, Romanian parts would be counted as local contents. Prior to its EU membership, Morocco had been importing parts from Romania, using these parts to assemble CBUs in Morocco, and attempting to export these CBUs to Europe. Since these Romanian parts did not qualify as local contents of the EU, the Logan did not have enough European parts to satisfy the rules of origin requirements for CBUs imported into the EU and was subject to the 10% duty rate on imported CBUs. However, with Romania's accession into the EU, the Romanian parts could qualify as European parts, and the Logan could satisfy the rules of origin to achieve a duty-free rate on CBU imports from Morocco to Europe. While Logan was introduced as a product for developing countries, the car was unexpectedly successful in Europe. Making use of the network of trade agreements, Renault was able to make use of the Morocco plant to fill European demands without having to pay hefty duties.

Figure 10.3 illustrates the supply network of the Logan car and the resulting duty-free flows as a result of Romania's accession into the EU.

This case example shows that, to fully capture the impact of regional trade agreements on customs and duties, it is *not* sufficient for us to simply look at the custom duty rate of a particular product from one country to another. In fact, we need to first examine the complete bill of materials of the product, and then the trade agreements

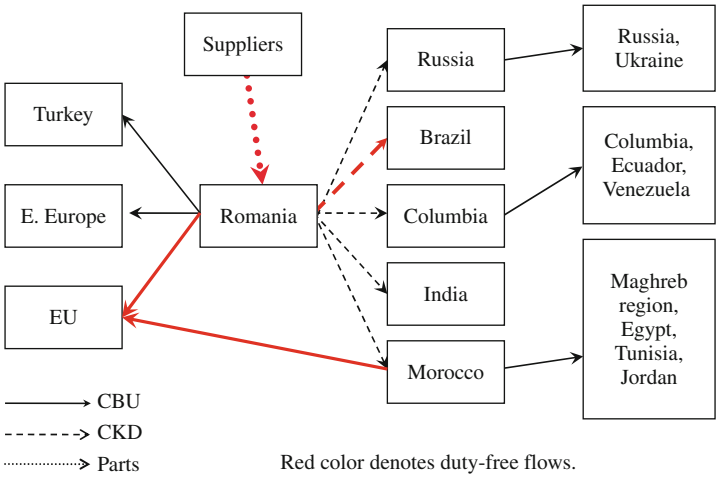


Fig. 10.3 Supply network of Logan car

of the components and the product among all the trading countries. For example, if Country A supplies CKD parts to Country B, which then supplies CBU to Country C, the trade agreements between Country A and B, Country B and C, and Country A and C must all be considered. Just as we need to consider the supply network, we now have to consider the network of trade agreements.

10.2.2 Trade Process Uncertainties and Risks

The complex trade agreements, regulations, and requirements set up by countries for trade could give rise to one big challenge to companies involved in trade. Even if one was able to figure out how to design the supply chain network to take advantage of the existing trade agreements and regulations, there is still a very high degree of risk that these agreements and regulations would change over time.

To illustrate, we return to the Logan car example of Renault. To reach the Egyptian market, Renault relied on the opportunities available by using Morocco as a trading center. As it was possible to import CKD parts into Morocco at a duty-free rate, Renault could import European parts into Morocco at a 0% duty and assemble the vehicles in that country. Renault could then export CBUs to Egypt from Morocco and obtain a 0% duty rate on these imports, a benefit of the free trade agreement between Morocco and Egypt. It was not possible, however, to export CBUs from Europe to Morocco and then export these to Egypt, as the CBU import rate into Morocco from the EU was 25% by the end of 2006. So using Morocco as the place to assemble the CKUs into CBUs for the Egyptian market seemed to be a smart move. However, such a design may not be optimal in 2009 as the duty rates on CBUs between Europe and Egypt were expected to decrease according to the Pan Euro Med protocol of origin. By 2019, a 0% duty rate for exports of CBUs from Europe to Egypt could be possible. Morocco may not be the optimal final assembly site then.

Apparel manufacturers have learnt first hand the uncertainties in trade agreements. When China entered WTO, the apparel quota for Chinese apparel products entering the United States was supposed to be phased out. Some manufacturers have closed down factories in other parts of the world, since these factories were not as efficient as the ones in China, in anticipation of the phasing out of quota from China. Of course, we soon learnt that the quota was not to be lifted totally. Some of these manufacturers were caught off guard.

10.2.3 Postponement Design

High technology products with a modular product structure can postpone some of the assembly processes to multiple global distribution points instead of integrating the complete product at the factory. Distribution points are much closer to the

customers, and so by allowing them to perform some of the final assembly processes, the point of differentiation of the product into multiple end-products can be deferred. Defining what is to be assembled in the factory, and what is to be assembled in distribution, is termed the *postponement boundary* problem. The labor cost rates, productivity, and customs and duty rates in the countries in which the DCs reside can be very different from those of the factory. These differences can have a significant impact on determining the best postponement boundary of a product.

Consider Hewlett Packard's (HP) workstation businesses in the late 1990s (see [10]). At the time, HP manufactured the workstation in two factories: one in the United States and the other in Germany. The factories distributed partially completed products to distribution centers in Europe and Asia as well as to a reseller network. This particular business also worked with six major resellers in the North American market and five major European resellers. The US and German factories also built fully configured systems for direct shipments to customers in North America and Europe, respectively, effectively serving as integrated factories and distribution centers.

When HP planned to introduce a new line of product with a modular design, it considered postponing some of the computer configuration processes to its DCs and even to its resellers. The two HP factories would continue to serve as their own distribution centers for their regions, which accounted for about 60% of all orders. For the rest of the orders, the postponement boundary problem would amount to defining the steps that were to be performed at the factories and those which were to be performed at the DCs.

The workstation product was sold in developed countries with high labor costs and moderate or high customs and duties levied on the product, such as Japan and parts of Europe that are outside of the EU; and in developing countries with lower labor costs but very high customs and duties levied on the product, such as Korea and Eastern Europe.

Figure 10.4 displays the total annual costs and the cost components for different postponement boundaries—the factory could build the complete product; or the product without storage and memory; or without storage, memory, and graphic boards; or without storage, memory, graphic boards, and processor; or without all the above plus the backplane; or delegate all key modules to be assembled at the DC. The bulk of the total costs were materials, and to highlight the cost differentials, we have chosen to show only the differential material and processing costs for each alternative relative to the least cost alternative for those particular costs.

The analysis showed that the best alternative was to assemble the chassis, power supply, and backplane assembly in the factory; this meant postponing the remaining steps, starting with the processor board, to the distribution centers and resellers. The U-shape result for the costs of the various options clearly indicated that the extreme options, building to order at the postponement sites and stocking fully configured units at the factories, were not cost effective.

As expected, inventory was the primary driver of the product configuration point; its effect followed the trade-off between the parts inventory created by postponing the activities and the stock of configured product. As the product content at the

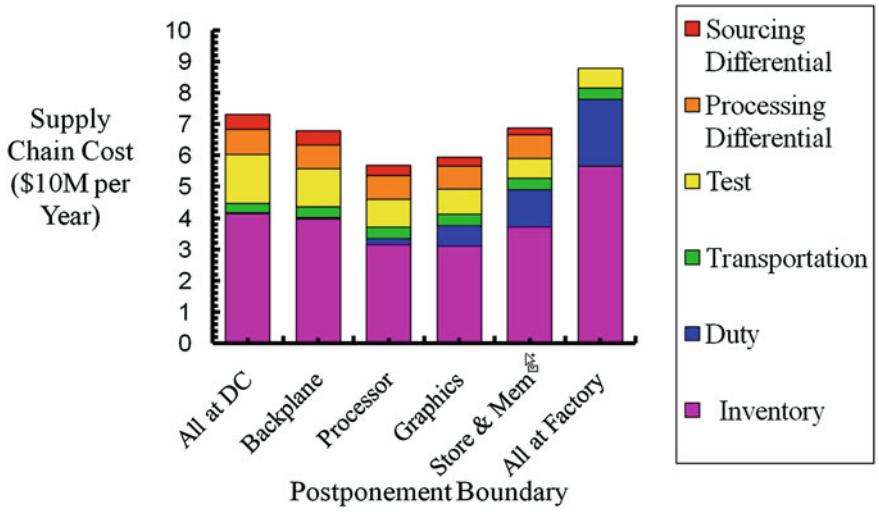


Fig. 10.4 Postponement boundary analysis

factories increased, the cost due to customs and duties also increased, since the configured product shipped from the factory to the DCs increased in its dutiable value. In addition, the duty rate applied to a product can also change depending on whether the product contained no processor, a processor, or a processor plus memory. This application case shows that, in a global supply chain, customs and duties can constitute a major cost driver in evaluating the postponement boundaries. Freight and processing costs did not factor as heavily in this particular example. The chassis accounted for the majority of the weight of the product, so there was little difference in freight expense between the different alternatives (we assumed high percentages of inbound surface transportation and outbound air transportation). The fixed costs of adding postponement capabilities to the distribution network were relatively insignificant. Because these other factors had so little effect, inventory and customs and duties made the difference in every scenario.

Hence, trade agreements and the resulting customs and duties affect the design of the supply chain, as well as the design of the postponement boundary.

10.3 Improving Global Trade Processes in Supply Chains

10.3.1 Logistics Efficiency and Bilateral Trade

As noted, the times and costs in global cross-border trade processes affect the total landed costs, which affect companies' sourcing decisions. Consequently, the efficiency of global trade processes could impact trade flows between two countries. The significance of this effect has been established by Hausman et al. [6].

Economists have long attempted to explain the variations of bilateral trades among nations by examining measures such as distances, the GDPs, and institutional quality factors (such as corruption and infrastructures). The gravity model has been a common means to perform statistical studies to examine the contributions of these factors toward explaining bilateral trade (see the review in [6]). Distance as a measure could certainly serve as a surrogate for the friction between two trading partners, negatively impacting trade. But in practice, this is only one of many factors that serve as frictions of cross-border trade flows. There are many process steps involved in cross-border trade, e.g., the times and costs required to make declaration, waiting for containers to be loaded on ships, customs clearance at both the exporting and the importing countries, the transportation time, inspection times at ports, and the times waiting for local transportation companies to bring goods to the final destination. There are also variances of these times that could also result in added frictions.

The Hausman et al. [6] study collected logistics performance metrics on some key cross-border trade flows across 80 countries, based on container flows of three key types of products: textile yarn, fabrics, made-up articles; apparel and clothing accessories; and coffee, tea, cocoa, spices, and manufactures thereof. These metrics were used in statistical analysis of an augmented gravity model, and the result was that cross-border global trade efficiencies (or the lack of) significantly impacted bilateral trade. While the best gravity model was able to explain about 66% of the adjusted R^2 of bilateral trade variation, the addition of cross-border logistics metrics improved the R^2 to 72%. As expected, the average process time to cross borders would negatively impact trade, but the variation of process times (the study used the difference between the maximum time and the average time as a surrogate for variation) also negatively impact trade. Hence, it is important for governments and companies to work on reducing both the mean and variance of times and costs of cross-border trade processes. Figure 10.5 shows the results of the augmented model in Hausman et al. (2009)

Hausman et al. [6] described some implications from the results of Figure 10.5. For example, the results can be used to see the benefits from reducing total process times through deregulating transportation, expanding ports to increase capacity, and promoting the growth of the third-party logistics industry to allow more consolidation of cargo flows. Trade-related processing time and cost can also be improved by re-engineering processes to eliminate unnecessary steps and streamline others (such as by introducing more parallel processing rather than sequential processing), introducing advanced information technologies (such as electronic customs clearance and documentation flows), using data mining and screening methods to identify only high-risk containers for security inspections, and adopting advanced scanning technologies to shorten cargo inspection times.

The model result can also be used to calculate the elasticity of the key logistics metrics to bilateral trade [6]. Let

$S(i, j)$ = value of bilateral trade from country i to country j;
$d(i, j)$ = distance from country i to country j;
$T(i, j)$ = average total time (transport and trade-related processing) from i to j;

Independent variable	*Coefficient*	*T-statistic*
Log of exporter's GDP	1.265	72.57
Log of importer's GDP	0.956	54.17
Log of distance	–1.390	–39.02
Exporter's Corruption Perception Index	0.188	10.82
Importer's Corruption Perception Index	0.134	6.27
Regional trade agreement dummy variable	0.343	4.73
Log of average time for all procedures	–0.373	–5.24
Log of total cost of procedures	–0.492	–10.68
Log of Maximum time-Average time	–0.236	–4.28

Adjusted R-squared: 0.716; Observations: 5149: F-statistic: 1287
Dependent variable is total bilateral exports (in logs) in 2003 or latest year available. Corruption Perception Index is for 2004, from Transparency International. OLS estimates; constant term not shown

Fig. 10.5 Augmented gravity model

$C(i, j) =$ total processing cost from country i to country j;
$\sigma(i, j) =$ maximum time minus average time from country i to country j.

Here, the (i, j) term in the variables can be suppressed without loss of generality. Figure 10.5 shows that

$$\log S = K' - 1.390 \log d - 0.373 \log T - 0.492 \log C - 0.236 \log \sigma,$$

where K' is a constant representing the non-logistic independent variables. It is easy to see that -1.390, -0.373, -0.492, and -0.236 represent the elasticities of the logistics metrics in bilateral trade. Thus a 1% reduction in the "distance" measure would be associated with an increase of 1.39% in bilateral trade. Regarding processing time, a 1% reduction would be associated with a 0.37% increase in bilateral trade. Similarly, a 1% reduction in the total trade-related processing cost would be associated with a 0.49% increase in bilateral trade, while a 1% reduction in the variability measure (maximum time−average time) would be associated with a 0.24% increase in bilateral trade.

10.3.2 Cross-Border Processes for Supply Chain Security

After September 11, 2001, the security of a supply chain has become a major concern to the public and the private sector. In particular, the ocean segment of a supply chain is most vulnerable to security threats, as more than 90% of world trade involves containers aboard ships [4]. The US government, in particular, has

been concerned with the threat of terrorism and the potential of having weapons of mass destruction (WMD) in materials flowing through a supply chain. WMD can result in significant loss in human lives, destruction of infrastructure, and erosion of public and business confidence. Ultimately, global trade and prosperity are threatened.

On the other hand, the private sector is concerned about the costs of assuring security and the potential disruptions associated with real or potential terrorist acts. Governments and industry have responded with proposals, such as increased information exchange among trading partners, ports, shipping companies, and the governments; and heightened inspection and scrutiny of the goods flowing through a supply chain. Increased inspection at the destination ports as a way to assure security can add tremendous cost, delays, and uncertainties in the supply chain.

US Customs has also launched the Container Security Initiative (CSI) and the Customs-Trade Partnership Against Terrorism (C-TPAT) in January and April of 2002, respectively. The C-TPAT program involves multiple countries and promotes the use of best security practices. Shippers and carriers that certify the use of best security practices are given expedited processing at US ports of entry. Manufacturers, importers, carriers, and third-party logistics service providers can all participate by completing detailed questionnaires and self-appraisals of their supply chain security practices, while Customs would perform periodic audits and verifications of such practices.

Another proposal was the Smart and Secure Tradelane (SST) initiative. This initiative has some similarities to total quality control (see [14]), which calls for having quality inputs and tight process control to assure quality, instead of relying on final inspection. Hence, rather than relying on inspecting containers arriving at the destination ports, we would focus on having containers inspected at the source and using technologies to monitor the transportation process to assure the integrity of the containers. Any tampering of the containers during the journey would have to be detected. To do this effectively, we do need to use modern technologies. One promising technology is the use of electronic cargo seals and sensors (smart containers). Such an initiative is not free, and so proper quantification of its benefits is crucial for general adoption.

The SST process starts with the identification of personnel, cargo, and transportation information about the container and its contents at the point of origin. This is followed by providing real-time supply chain security and management information to partners involved in the end-to-end shipment, through integrating data from Active-RFID (radio frequency identification) tags and intrusion-detection sensors attached to the containers. The RFID tags are read by stationary and mobile readers at key nodes.

Simple models can be developed to assess the benefits of SST (see [13]). Let

p = the inspection rate of containers arriving at a destination port;
x = transit lead time in days, a random variable;
y = inspection dwell time in days, a random variable;
T = total lead time in days.

Note that

$$E(T) = E(x) + pE(y) \quad \text{and} \quad \text{Var}(T) = \text{Var}(x) + p\,\text{Var}(y) + p(1-p)[E(y)]^2.$$

μ = mean daily demand of a product;
σ = standard deviation of the daily demand of the product;
R = inter-replenishment time in days for the DC;
k = safety stock factor;
p' = new inspection rate under SST;
$1-\theta$ = percentage reduction of the transit time variance as a result of SST.

Hence, the new transit time variance under SST is given by $\theta\,\text{Var}(x)$.

Without SST, i.e., in the current process, the safety stock is given by (see, for example, [15]):

$$S_0 = k\sqrt{\mu^2\,\text{Var}(T) + \sigma^2 E(T+R)}.$$

With SST, we have advanced information about the lead time statistics and so could adjust the safety stock based on the knowledge of whether inspection is needed or not. The resulting expected safety stock is

$$S_1 = k\Big\{p'\sqrt{\mu^2\,[\theta\,\text{Var}(x) + \text{Var}(y)] + \sigma^2\,[E(x) + E(y) + R]} + (1-p')\sqrt{\mu^2\theta\,\text{Var}(x) + \sigma^2\,[E(x) + R]}\Big\}.$$

Lee and Whang [13] showed that $S_1 \le S_0$, providing one benefit of SST.

10.3.3 IT-Enabled Global Trade Management for Efficient Trade Process

Using advanced information technologies (IT) on some process steps in cross-border processes to assure supply chain security is one way in which we can improve the trade process and gain some benefits. But there are many other process steps that could also benefit from process improvements through the use of IT. IT can of course potentially reduce the mean and variance of the lead time in a process step (through direct work savings and reduction in errors and reworks). But it can also enable, in some cases, parallel processing of some process steps instead of sequential. It can allow for some re-sequencing of the process steps that could lead to overall savings. Finally, it is also possible that some process steps can be eliminated (e.g., if IT results in a process with zero defects, then another subsequent step for the purpose of checking and verification can be eliminated). Hence, investment in IT can be a powerful way to improve cross-border trade processes, which would then lead to supply chain performance improvements.

To do a complete analysis of the potential process improvements, we need to (1) characterize all the process steps involved in trade flow, as well as their precedence

relationships; (2) estimate the current duration and cost for each of the process step; (3) estimate new process flow with IT fully implemented, and the resulting duration and cost for each of the process step; and (4) given these changes, quantify the benefits to the exporters, importers, and other intermediaries involved in trade processes. Teamed with TradeBeam, a leading IT provider of trade processes, Stanford University has developed the Stanford Trade Process Model for the purpose of performing such an analysis (see [7]).

The trade process is extremely complex. This is partly due to the proliferation of regional trade agreements described earlier. Compliance to these agreements requires extensive documentation, tracking, and verification, all of which become part of the cross-border processes. The term Global Trade Management (GTM) refers to the processes required to support cross-border transactions between importers, exporters, their trading partners, and governments. GTM encompasses network planning, sourcing, order collaboration, compliance with government regulations, transportation, inventory, and warehousing management, as well as financial settlement. GTM can be performed manually, or in a highly automated fashion, and with poor or efficient processes. Information Technology-Enabled Global Trade Management (IT-GTM) is the set of information technologies and software solutions that can be used by companies to streamline and perform their global trading processes. They can include automation of export and import management and compliance, electronic integration with trading partners, trade financing, and trade content management.

The Stanford Trade Process Model is focused on apparel trade from China to the United States and is based on extensive interviews and data collection with trade experts and companies involved in such trades [7]. It involves over 100 process steps that cover broadly the following processes (Figure 10.6):

Fig. 10.6 Stanford trade process model

Pre-export: initiation of the global trade process, e.g., import screening, price negotiation price, contract and payment terms, creation of purchase/sales orders, and export screening;

Transport arrangement and export declaration: preparation for exportation, including arrangement of transportation carriers, obtaining approval from inspection agencies, export declaration, and preparation and transmission of security filings to US Customs and Border Protection;

Transport and import declaration: ocean or air transport of the goods, generation and submission of import documents, and import customs clearance;

Post-import customs clearance and payment: inland delivery from the border to the importer's site, receipt of goods, review of landed cost, settling payment with the forwarder, broker and exporter, and filing for foreign exchange verification and tax refund if applicable.

IT-Enabled GTM could result in direct process improvements or process reengineering which have tremendous benefits (Figure 10.7).

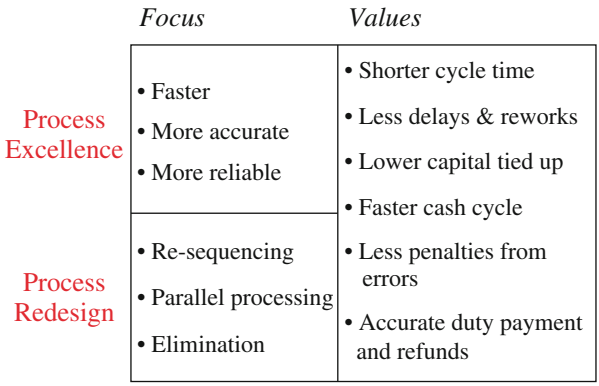


Fig. 10.7 Values of IT-enabled innovations

To quantify the benefits of such improvements, one needs to develop models to capture the benefits in the form of inventory savings, savings in financing costs, speeding up tax rebates, reduction in expedite costs, reduction in fines, logistics cost savings, labor cost savings, potential reduction in procurement, markdown and lost revenues for importers, and customs savings due to accurate classification of products. Intermediaries (such as banks, freight forwarders, and other service providers) can also benefit through workload reduction and reduced cost of receivables financing. The benefits to exporters, importers, and intermediaries have to be modeled separately. The analysis of the apparel trade from China to United States shows that the value of IT-Enabled GTM can be significant (see Figure 10.8).

To illustrate, Hausman et al. [7] show that, for exporters, the order to receipt cycle could reduce from 104 to 68 days; the number of days outstanding could drop from 42 to 30; and the manufacture to invoice cycle could be shortened from 45 to

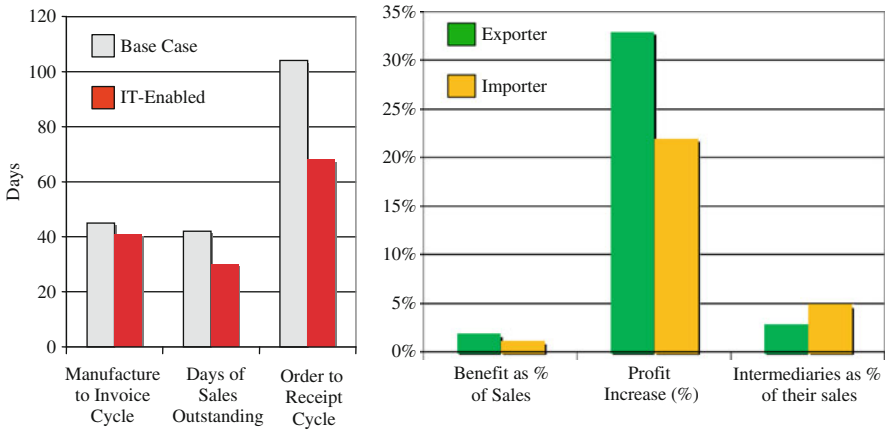


Fig. 10.8 IT-enabled benefits

41 days. The annual benefits of IT-Enabled GTM are 1.7 and 1.4% of annual sales for the exporter and importer, respectively. Assuming net profit is approximately 6% of sales for both exporters and importers, these improvements represent a 28% increase in annual profit for exporters and a 23% increase in annual profit for importers. Intermediaries for exporters and importers could also realize benefits amounting to 3 and 5.5% of annual sales, respectively.

10.3.4 Empirical Analysis of Trade Processes

Given the importance of a deep understanding of the trade processes, it is crucial to have a solid picture of the performances of the trade process steps empirically. The Stanford Trade Process model [7] made use of interviews and questionnaires sent to trade experts and practitioners for data collection on the performances of the process steps. A more concrete approach is to obtain the information from real data. This was the approach undertaken by Lee and Lim [9]. The study focused on cross-border processes between Shenzhen, China, and Hong Kong. It detailed all the process steps involved in clearing customs, transporting goods, and other logistics processes. There were heavy cross-border traffic between China and Hong Kong, but the study was based on trade flows related to the Outward Processing Arrangement (OPA). Under OPA, some core apparel products made in China would be shipped to Hong Kong for some assembly steps, then back to China for some finishing steps, after which the products would pass through the Hong Kong port and be exported to the United States with the origin declared as Hong Kong, thereby avoiding the quota constraints imposed by the US government on Chinese imports. OPA is a legitimate process, provided that the right job content is carried out (and verifiable) in Hong Kong. Hence, the products would have crossed the China–Hong Kong border three times (first China to HK, then HK to China, and then China to HK).

To collect real data, the researchers installed GPS and RFIDs on a sample of trucks and had readers mounted at some key choke points along the China–Hong Kong border. Figure 10.9 shows the points in which readers were installed so that the movements of the trucks were tracked. Consequently, the actual times required to go through all the cross-border process steps could be recorded, leading to some very concrete estimates.

- Total: 17 points along the cross-border route between China and Hong Kong
- Lok Ma Chau yellow bus stop: A1-A4
- Lok Ma Chau Hong Kong Customs: B1-B5
- Huanggang customs: C1
- Pedestrian bridge at Huanggang customs: D1
- Pedestrian bridge besides Guangyin building: E1
- Riverside besides Guangyin building: F1
- Southern point of crossroad outside Huangyuyuan north gate: G1
- Customs truck north exit: H1
- Truck park north entrance: H1
- Huanggang Customs truck entrance: J1

Fig. 10.9 Data collection for sample points

		HK			SZ		
		Queue	Process	Inspection	Queue	Process	Inspection
HK to SZ (5,258 trips)	Rate			17%			14%
	Mean	3.00	1.25	6.54	18.93	3.03	20.81
	Std Dev	5.17	2.38	14.36	39.46	7.51	30.48
SZ to HK (4,662 trips)	Rate			8%			19%
	Mean	1.03	1.01	9.44	3.17	3.08	58.62
	Std Dev	2.15	2.00	23.57	5.74	5.32	70.61

Fig. 10.10 Cross-border cycle times (hours)

After extensive data collection, the study was able to have very accurate estimate of the inspection rate performed by customs office, the means and standard deviations of the queueing and process times in crossing the border on the Hong Kong (HK) side as well as the Shenzhen (SZ), China, side. Figure 10.10 shows the statistics.

Hence, using GPS and RFID can be one way to get real-life data on the trade processes.

10.4 Concluding Remarks

As a result of increased globalization of industrial supply chains, effective supply chain management requires sound alignment with the global trade processes. I have discussed how the design of the global supply chain and the determination of right level of postponement are both tied intimately to the prevailing network of trade agreements, regulations, and local requirements of the countries in which the company is operating in. Moreover, the dynamic changes and uncertainties of these agreements and requirements must be anticipated.

In addition, the complexity of the cross-border trade processes results in uncertainties in the lead time and costs involved in global trade, which naturally forms part of the consideration of global sourcing, and the resulting safety stocks or other hedging decisions. Governments, exporters, importers, carriers, and other service providers have to work together to reduce the logistics frictions involved in the global trade processes. The benefits accrue not only to the exporters, importers, and the intermediaries, but ultimately, they could foster bilateral trade. The only way to reduce the frictions is to gain a deep understanding of the detailed process steps involved, to improve upon it by IT, and potentially re-engineer the processes. But the payoffs to such investments can be huge.

The inter-relationships between global trade processes and supply chain management form a fertile ground of research. I hope that the above discussion can stimulate ideas for this purpose.

Acknowledgment

This chapter draws upon past and ongoing research that I have done with colleagues like Professors Chung-Yee Lee, Warren Hausman, Lingxiu Dong, Seungjin Whang, and Morris Cohen.

References

1. Arntzen BC, Brown GG, Harrison TP, Trafton LL (1995) Global supply chain management at digital equipment corporation. Interfaces 25(1):69–93.

2. Cohen MA, Lee HL (1988) Strategic analysis of integrated production-distribution systems: Models and methods. Operations Research 36(2):216–228.
3. Cohen MA, Lee HL (1989) Resource deployment analysis of global manufacturing and distribution networks, J Manufacturing and Operation Management 2(2):81–104.
4. Cuneo EC (2003) Safe at sea. Information Week (April 7)
5. Geoffrion AM, Graves GW (1974) Multicommodity distribution system design by benders decomposition, Management Science 20(5): 822–844 (also reprinted in this volume)
6. Hausman WH, Lee HL, Subramanian U (2009) The impact of logistics performance on trade. Production and Operations Management (under review)
7. Hausman WH, Lee HL, Napier G, Thompson A, Zhang K (2010) A process analysis of global trade management: An inductive approach. Journal of Supply Chain Management (to appear)
8. Hoyt D (2007) Crocs: Revolutionizing an industry supply chain model for competitive advantage. Stanford Graduate School of Business case GS-57, May 9
9. Lee CY, Lim A (2008) RFID cross-border project: Process enhancement feasibility study. Hong Kong University of Science and Technology
10. Lee HL, Dong L (2009) Postponement boundary for global supply chain efficiency. Working paper
11. Lee HL, Shao M (2009) European recycling platform: Promoting competition in e-waste recycling. Stanford Graduate School of Business case GS-67, August 18
12. Lee HL, Silverman A (2008) Renault's Logan car: Managing customs duties for a global product. Stanford Graduate School of Business case GS-62, April 29
13. Lee HL, Whang S (2005) Higher supply chain security at lower cost: Lessons from total quality management, International Journal Production Economics 96:289–300.
14. Lee HL, Wolfe M (2003) Supply chain security without tears. Supply Chain Management Review 7(1)(Jan/Feb):12–20
15. Silver E, Pyke D, Peterson R (1998) Inventory management and production planning & scheduling. 3rd ed. Wiley, New York, NY

Chapter 11
Sustainable Globally Integrated Enterprise (GIE)

Grace Lin, Ko-Yang Wang

Abstract In this chapter, we present the globally integrated enterprise (GIE) as an emerging business model with strong implications for how companies run and operate their global supply-and-demand chains. The GIE shifts the focus from an efficiency-driven model to a value-driven one which leverages and integrates global capabilities to deliver value speedily, seamlessly, and in a flexible way, while maximizing profits. A GIE is a complex organization that faces many challenges. The evolution of the supply chain in the last 20 years has paved the way for the Operation Research (OR)-enabled Sense-and-Respond Value Net that supports today's GIE needs. We present a GIE case study of a business transformation journey. We then describe the next steps for GIEs to become more socially, economically, and environmentally responsible through the use of OR, business analytics, and IT.

11.1 Introduction

> A Globally Integrated Enterprise (GIE) is an open, modular organization that is integrated into the fabric of the networked economy and operates under a business model that makes economic sense in the new global landscape [14].

Several fundamental changes in the last 20 years have caused multinational companies to rethink their approach:

1. The breakdown of economic nationalism caused trade/investment barriers to recede, accelerating the globalization trend.
2. Advances in technology and open standards have significantly improved the speed and reduced the cost of global communication.
3. Geopolitical changes opened up new markets and skill pools which had been unexplored by multinational corporations.

Grace Lin, Ko-Yang Wang
World Resource Optimization Inc., Chappaqua, NY, USA; Global Business Services, IBM, Armonk, NY, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_11,

These changes have caused companies to re-evaluate how they manage their business since where and how business value is created in this new environment is evolving. For example, sharing work across country or continent borders becomes possible, and outsourcing and global operations seem much more appealing. Many US and European companies have moved or outsourced some or all of their manufacturing or services to Asia, South America, and East Europe, with increasing speed. The multinational companies' traditional approach of replicating themselves and building plants locally while maintaining some key corporate functions such as R&D and product design in their home countries is no longer sustainable. They need to create flatter, more efficient operating models while building new innovative capabilities globally to drive profitable growth. This new model has implications for how companies run and operate themselves and their global supply chains. Fundamentally, the focus has shifted from simply managing the supply chain for greater efficiency to leveraging it to drive revenue, profit, and customer satisfaction. In the 2008 IBM Global CEO study, CEOs indicated that they were embracing the global integration and unpredictability as the new routine [22]. They were also anticipating the need for their business to respond seamlessly and globally with unprecedented speed and flexibility. In his 2006 article in "Foreign Affairs," Sam Palmisano, IBM's CEO, coined the term "Globally Integrated Enterprise" (GIE) for this emerging business model [45].

In his article, Mr. Palmisano pointed out four major challenges for a GIE: (1) securing a supply of high-value skills; (2) creating sensible worldwide regulation of intellectual property; (3) determining how to maintain trust in enterprises based on increasingly distributed business models; and (4) managing requirements for long-term vision and continuous investment from business leaders. Recognizing the scale of these challenges, Mr. Palmisano called for the leaders in business, government, education, and civil society to learn the emerging dynamics of GIEs and to help GIEs mature in ways that would contribute to social, economic, and environmental progress around the planet. Two years later, in his 2008 speech to the Council on Foreign Relations, Mr. Palmisano discussed "A Smarter Planet: The Next Leadership Agenda" [46]. He described the IBM vision for a Smarter Planet and the way in which the world is becoming instrumented, interconnected, and intelligent. He laid out visionary scenarios that lead the way to transforming companies into GIEs and also pointed out a new direction toward sustainability, asking IBM, business, and civil leaders to jointly work on specific solutions.

The key motivations for the multinational companies to go global have remained the same: to improve revenue and profits by entering new markets, reducing production costs, and seeking skilled workers at low costs [24]. However, in the face of accelerating change brought about by globalization, technology advances, standardization, competition, and geopolitical evolution, as well as the skills evolution of both developing and developed countries, the operational model of the multinational companies is undergoing fundamental changes structurally, operationally, and culturally and at an unprecedented pace. The benefits of a well-run GIE are obvious: With the support of global skills and communication, the GIEs are able to strategically place their operations anywhere in the world that offers the lowest cost or the

best strategic value. However, transforming companies into well-run GIEs is not an easy task. They require fundamentally different approaches to production, distribution, workforce management, product design, and risk management, etc. Any misstep or miscalculation can cause significant cost and damage to the company.

In the last few years, more and more companies have started to examine their sustainability. For example, IBM and its clients' Smarter Planet efforts [Smarter Planet, Web], Dubai's SmartCity [SmartCity, Web], US government and utility companies' Smart Meters, and Intelligent Grids initiates [48]. Thus, we define the sustainability of GIEs: *The ability to improve business performance while reconciling the company's needs with those of its supporting ecosystems from an environmental, social, and economic perspective*. We believe that (1) the emerging GIE trend is accelerating and it is important for enterprises and society to embrace it and focus on maximizing the positive impact of GIEs on business performance while minimizing the negative impact of GIEs on the economy, society, and the environment and (2) sustainability is a critical success factor for the GIEs. This means that enterprises have a social responsibility to ensure that their pursuit of maximizing profits and minimizing costs has a positive impact on the sustainability of the economy, society, and environment. Also, the transformation of GIEs should improve their own sustainability, i.e., their ability to manage, survive, and even prosper while facing unexpected environmental changes, disasters, or disruptive events.

In this chapter, we focus on the supply-chain and value-net aspects of the GIEs and their sustainability. We will examine the recent advances in supply-chain management (SCM) and information technology and their critical role in GIEs. In Section 11.2, we briefly touch on the key challenges that today's GIEs face. In Section 11.3, we discuss the evolution of SCM and how it has driven the major changes seen in today's GIEs. We will also discuss using the OR-based adaptive Sense-and-Respond Value Net to better enable GIE. In Section 11.4, we review a case study and examine some best practices in improving business performance and the process of transforming a company into a GIE. Finally, in Section 11.5, we discuss the characteristics of effective GIEs and how these companies can become more sustainable and socially responsible by leveraging advanced SCM solutions.

11.2 An Overview of GIEs and the Challenges they Face

> "The crisis in our financial markets has jolted us awake to the realities and dangers of highly complex global systems. But in truth, the first decade of the 21st century has been a series of wake-up calls with a single subject: The reality of global integration." Sam Palmisano [45]

Companies started to move or outsource their operations abroad in the late 1980s but the trend toward GIEs has accelerated in the last 10 years. With the advances in technology, we see communication and collaboration becoming easier and less costly and location, distances, and geographic borders becoming less relevant. The decisions about where, whom, and how products and services are made or provided

are driven less by the "where" and more by cost, skills/knowledge, and even eco-political considerations. We are not only seeing companies moving labor-intensive manufacturing to lower cost countries such as China, India, or Brazil but also witnessing components of skill-intensive products/parts/services being moved in the same manner and integrated back into the corporate processes on a global scale. To understand these changes and how to transform enterprises into successful GIEs, we will first examine the key challenges that GIEs face today.

The financial meltdown of October 2008 and the subsequent collapse or near collapse of the financial, housing, construction, automotive, and many other industries did not happen overnight. The problems were years in the making. However, for most enterprises, the sudden realization of the realities and the dangers of venerable business models and financial stability had executives scrambling to rethink their strategy and operational models and to seek new solutions. The scale and reach of the crisis and the speed with which some seemingly infallible companies crumbled, as well as the vulnerability of companies in general, surprised virtually everyone.

The crisis highlighted several key realities of today's business environment:

1. *We are all connected.* This close interdependence exposes enterprises to risks that they cannot totally control. Today's enterprises operate in a complex web of business relationships so that interdependence is deeply rooted in the fabric of the business model. Close collaboration with business partners is essential for performance improvement.
2. *It's a small world.* The financial crisis triggered by the US mortgage industry would have brought down almost all major financial institutions in the major financial markets had central governments not intervened and saved them. This crisis revealed the fragility of the business models of many enterprises (e.g. the auto industry's crisis triggered by consumers tightening up during the financial crisis). Sustainability should be a key focus of any enterprise.
3. *When the market changes faster than a company's ability to react to it, the company is in trouble.* Unfortunately, not every company can keep pace with the acceleration of market changes. For example, the major American auto companies that relied heavily on SUV and truck sales for profit found that consumers had changed their buying habits in the face of the great financial crisis in 2008. As the crisis deepened, auto sales slowed to half their size from the previous year. Although some companies recognized the peril of the market earlier (with the auto market slowdown beginning in early 2008), they were unable to adjust quickly enough which resulted in the auto industry bailout in 2009. This highlights the importance of agility and flexibility.

Another incident in the auto industry, the Toyota "sticky gas paddle and sudden acceleration" issue which continues to involve millions of car recalls in 2010 also demonstrates that a company is only as strong as the weakest link in its supply chain. Toyota, with its stellar reputation of reliability, stopped selling half of its cars overnight because of a faulty part from a weak link in its supply chain causing the recall. The problem was compounded by Toyota's initial slow response to a key crisis. The long-term damage to this widely admired company is immeasurable.

Today, enterprises face the following key challenges:

1. Labor is only one of the many costs of global operation. Relocating operations to low-cost areas also increases the risks of disruptive missteps due to increased complexity, communication, and logistics issues. Simply replicating existing operating models will not work well.
2. The rapid pace of market changes often renders business models obsolete before transformation is complete or becomes effective.
3. Technology advances are accelerating process automation and enterprise collaboration but many companies are confused by different incompatible technology standards in business modeling and process automation.
4. Information integration is a critical step in enabling intelligence business analysis but to integrate monolithic applications and clean up data is expensive and time consuming. Intelligence and business analysis too have to be explored.
5. Agility and flexibility are a reflection of the business model and operational process model. IT technologies are critical enablers but resistance to change often reduces or prevents a company's form being agile or flexible.
6. The social and economic impact of GIEs' "cherry picking" can have a profound impact on the communities/countries they abandon; and the backlash can impact the customer relationship/markets as well.

Enterprises can leverage experience, R&D, technology, and natural evolution in a holistic approach that will allow them to transform into GIEs, enhance their agility and flexibility to improve business performance, and, more importantly, become more sustainable.

11.3 The Evolution of Supply Chains and the Sense-and-Respond Value Net

Business and information technology advances in the last decade, particularly in business analytics, process modeling and automation, information integration, and business performance management, present new opportunities for enterprises to enhance their ability to compete. At the same time, converging social and technological trends are changing the nature of decision-making to create a more collaborative approach [37]. The evolution of the supply chain in the last 20 years has paved the way for the OR-enabled Sense-and-Respond Value Net to leverage these technologies to support today's GIE needs.

Over the last two decades, companies have evolved from the internal functional and process efficiency transformation toward collaborative and adaptive GIEs. SCM has been vital in many companies' transformation success such as Toyota [31], Nokia [16], Dell [17], Amazon, and IBM . In this section, we will discuss the supply-chain evolution and lessons learned. We will also describe the adaptive Sense-and-Respond (S&R) Value-Net model and its applicability for GIEs.

In the early 1990s, Enterprise Resources Planning (ERP) was adopted by many companies as a means for automation and improving transactional efficiency.

However, we have seen the top-down, ERP-based processes being stressed past their intended capabilities for transitional efficiency. A major issue with the ERP systems is their lack of flexibility and speed to support decision making throughout the internal and extended supply chain to meet changing business needs. By the mid-1990s, various Advanced Planning and Scheduling (APS) tools implemented with legacy and ERP systems were developed to support the optimization of supply chains during planning and execution cycles. Business Process Reengineering, Just-in-Time, and Lean Supply-Chain process implementation were also becoming major trends. In the late 1990s and early 2000s, the development of e-commerce and e-business tools offered Internet connectivity and some limited capability for supply-chain collaboration and near real-time information sharing (see Figure 11.1).

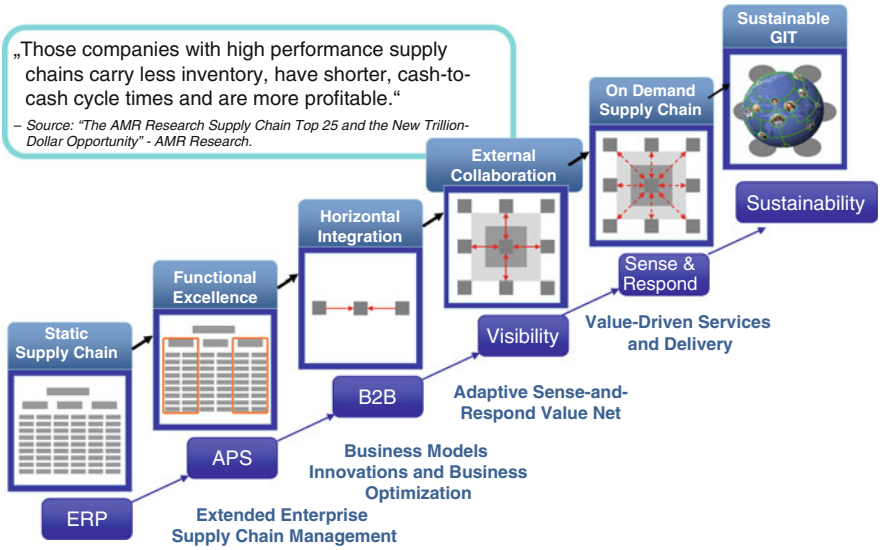


Fig. 11.1 The evolution of the enterprise—Enterprises have been evolving from functional and process efficiency to a collaborative and adaptive globally integrated enterprise (Source: IBM Sense-and-Respond Presentation, 2004)

However, despite the implementation of supply-chain management tools and Internet connectivity, the ROI of supply-chain management package implementation has constantly come under question. Based on interviews with senior executives from 25 firms, Forrester reported that companies overspent on supply-chain optimization packages and received diminished returns: 80% of the companies spent more time than expected and, on average, companies spent 74% over budget to implement supply-chain optimization tools [49]. It was also reported that product markdowns due to excess inventory jumped from 10% to 30% of total units sold while customer satisfaction with product availability plummeted [30].

Why did efficiency gains and automation fail to delivery business value? In the changing business environment, disruptive business and technical events can occur

any time and at every level. Major business disruptions and inefficiencies can be the result of the inability to handle these events quickly and intelligently due to (1) a lack of information visibility across internal and external supply chains; (2) insufficient partner collaboration; (3) a lack of customer intimacy; (4) an inability to leverage knowledge and manage uncertainty; and (5) a lack of flexibility in business processes, applications, and infrastructure. The just-in-time supply-chain model performed well in improving supply-chain efficiency and minimizing product defects. However, the model depends on the ability of its supply network to control their inventory and deliver parts "in time." If not, the process breaks down. Furthermore, local supply-chain optimization based on incomplete or disjointed information under rigid top-down planning models can not only result in sub-optimization but also cause significant adverse effects. Therefore, since early 2000, many supply-chain experts have vigorously started to explore new models that expand the supply-chain scope and allow more agility and flexibility. Studies of some of the more successful supply chains of the early 2000s revealed that supply-chain optimization depends on their ability to streamline operations while processing information intelligently and holistically, allowing quick, proactive, and effective responses to frequent changes in the market place. This includes understanding the needs of customers and the needs and capabilities of business partners and employees, as well as gathering relevant information to analyze risks and opportunities and gain situational awareness in changing environments [2, 21, 27–29, 35, 39, 51].

These studies concluded that the key to successful, adaptive organizations is to ensure continued focus on responsiveness and agility. This, however, cannot be achieved through technology implementation alone but by transformation into a business model supported by real-time business processes and performance management that will allow to quickly evaluate situations and determine how best to adjust business models, processes, applications, or partnerships to key issues and events. This is what we called the "Sense-and-Respond Value-Net Model" [39], what Hau Lee called the "Triple A Model," [29, 36] and what AMR called "Demand-Driven Supply Chain" [10].

AMR in particular defined the "demand-driven supply network" (DDSN) model that transforms a factory-oriented "push" set of activities to an innovation and supply capability driven by the demands of customers. AMR research has benchmarked business processes in detail [e.g., 16] and found a clear correlation between leadership in the use of demand-driven principles and tools and higher level financial metrics. Professor Hau Lee identified the three characteristics of successful supply chains based on supply-chain success stories including Wal-Mart, Dell, and Amazon: Agility, Adaptability, and Alignment. Lee concluded that to achieve a sustainable competitive advantage, a supply chain needs all three of these qualities simultaneously.

The Sense-and-Respond Value-Net model was introduced by Lin et al. at IBM in 2000. The objective was to build an open and adaptive framework to enable value-driven business optimization. A Sense-and-Respond Value Net was then a new paradigm that integrates real-time decision support, risk and resource management, supply-chain optimization, business processes automation, and partner

alliances in an integrated management system. Through its sensing, responding, and analyzing capabilities, a Sense-and-Respond enterprise monitors and evaluates real-time business performance and market conditions, aligns operations with strategy and customer requirements, proactively detects events, and engages value-net partners in collaborative decision making (see Figure 11.2). It could be viewed as a digital brain with sensors reaching all the way from a company's global value chain to the Internet world, blending business and IT to support value-net optimization in uncertain and dynamic environments [39].

In 2004 Lin et al. [36] presented a framework for S&R value-net transformation as well as a maturity model for identifying gaps and defining roadmaps. They defined the five areas needed to support the development and adoption of the Sense-and-Respond model: *(1) Adopting Sense-and-Respond Managerial and Technology Transformation with a focus on culture, (2) Support for Integration, Collaboration, and Security, (3) Information Intelligence, Analysis and Trustability of Data and Their Aggregated Impacts, (4) Modeling Uncertainty and Managing Performance, and (5) Support for Agent Systems and Distributed Decision Support.* In 2006, they further identified and studied the S&R technology enablers and concluded that most enabling technologies are actually available today (see Figure 11.3).

Within the last 10 years, the Sense-and-Respond model has been adopted by many companies and software vendors, as well as by United States and international defense agencies [9]. We will discuss IBM's success story in Section 11.4.

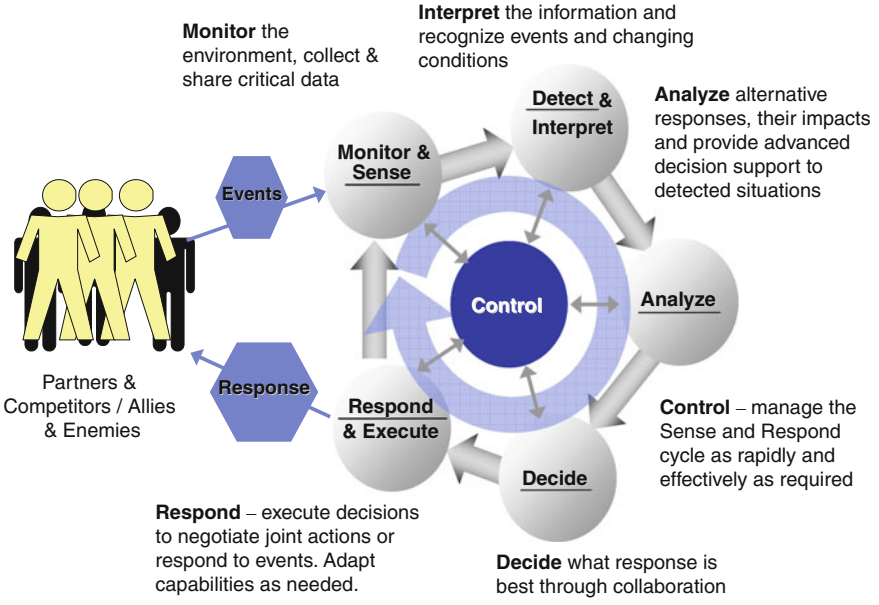


Fig. 11.2 The Sense-and-Respond operational model (Source: [36])

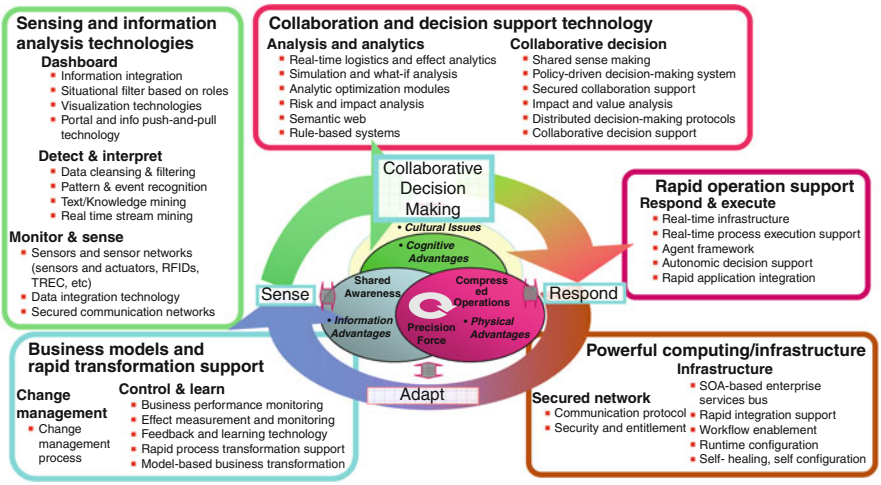


Fig. 11.3 Most key technologies needed for enabling Sense-and-Respond value nets are available today (Source: [7])

To summarize, we have seen the supply-chain transformation focus change from efficiency-driven automation, cost reduction, and streamlining of the supply-chain processes to information- and collaboration-driven extended supply-chain integration to value-driven adaptive Sense-and-Respond value net, which combine

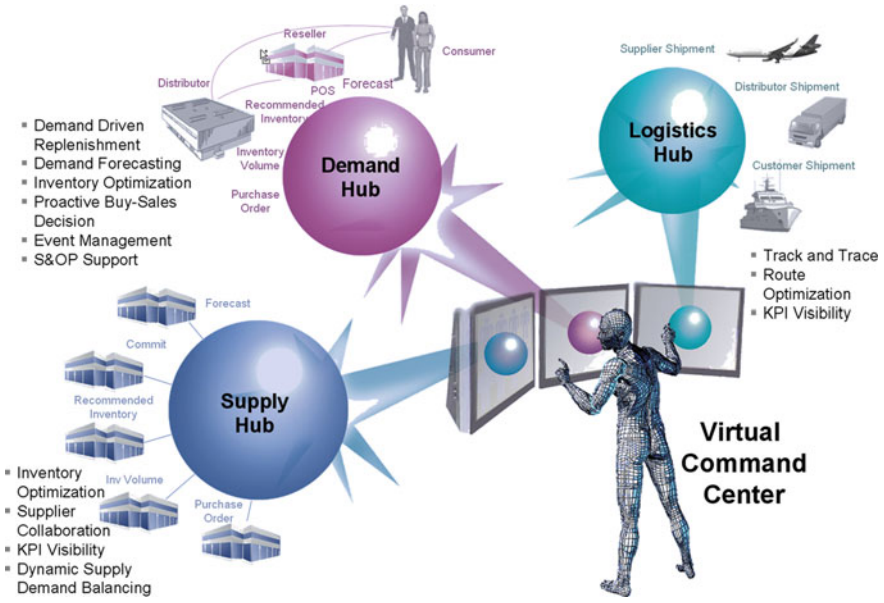


Fig. 11.4 VCC provides end-to-end visibility to enable value-net partners to collaboratively support chain performance (Source: IBM VCC Presentation, 2008)

information integration, adaptive process enablement, and business analytics to improve collaboration and the quality of decisions.

The next logical step for GIEs is to fully utilize S&R to integrate supply, demand, logistics, and other key business functions globally and to perform S&R culture transformation to become a true GIE.

IBM's recent 2008 CEO study of more than 1,000 C level executives found that the vast majority of companies are becoming globally integrated, with 75% actively entering new markets [22]. Of those, 84% plan to partner with local companies to become truly globally integrated. Companies that can master the enabling technologies shown in Figure 11.4 and integrate them into the fabric of their business to transform their business model to become more instrumented, connected, and intelligent will have significant competitive advantages.

Some success stories in GIE are already being reported, such as already mentioned Wal-Mart, Nokia, and IBM. In Section 11.4, we will discuss how analytics/OR was used in combination with business process modeling and innovative business models to support IBM's GIE Transformation.

11.4 A Case Study

> IBM has one of the largest and most complex supply chains in the world. Being a technology leader in the Industry, IBM needs pragmatic and powerful supply-chain technology to address its business challenges driven by rapidly changing business environments. Over the last 20 years, it has demonstrated a compelling story in business transformation and global integration.

IBM's transformation in the last 20 years makes a great case study for the Global Integrated Enterprise. The company's reengineering effort of the 1990s began out of necessity. From the late 1980s to the early 1990s, only several years after recording its record-high revenue, IBM suffered a remarkably rapid fall from grace due to its slow reaction to a market transformation from mainframe computing to a distributed computing model. Both its technology and its relationship with customers were viewed as antiquated. In 1993, a victim of its size, bureaucracy, insular culture, and the workstations and PCs that it had helped invent, the company lost a record $8.1 billion. At that time, IBM's cost structure was too high; the company was too decentralized; it stayed with an old strategy too long; and it had lost touch with both industry changes and its customers. IBM and its mainframe were dubbed dinosaurs and their imminent extinction was predicted [43]. With a pending plan to break up the company, IBM hired Lou Gerstner as CEO in 1993. Realizing that the real value IBM offered its customers was its ability to provide end-to-end solutions to business problems, Gerstner [18] reversed course and set a strategy to create a unified, integrated company. As part of this effort, IBM brought together its operations into a globally integrated supply-chain organization. It changed its manufacturing from build-to-plan to build-to-order. It started its services business and aligned its products and solutions to provide end-to-end solutions to their clients. The strategy paid

off. By the year 2000, IBM's net income had grown to $8 billion—a $16 billion turnaround from the dark days of 1993.

Beginning in 2002, IBM embarked on the second phase of its transformation journey. Its strategy was to become the showcase example of an on-demand business and innovation company. This transformation was no longer made critical by a burning platform and a struggle for survival but rather driven by a collective aspiration to turn a good company into a great company again. IBM continued to change its business model, its operations, its processes, and its culture to respond to the changing demands of globalization in the 2000s. In 2008, IBM posted excellent results despite an ailing global economy: the company had record revenue of $103 billion, a profit of $16 billion, earnings per share of $8.93, and a record cash flow of $15 billion, up almost $2 billion year to year.

IBM's Integrated Supply-Chain transformation is a cornerstone of IBM's transformation success over the past decade, coupled with the applications of advanced OR and BPM technology. The company has turned the management of its nearly $40 billion in annual spend into a disciplined application of services science, one that has produced billions of dollars in cost savings and contributed in a major way to IBM's steady improvement in earnings per share over the past several years.

A decade ago, IBM's supply chain was fragmented throughout the company in various business units and operating structures. Bringing together these organizations into a globally integrated supply chain can be complex and risky. The Integrated Supply-Chain (ISC) organization was created in 2002 as a single business unit, charged with making the company's supply chain a competitive advantage, i.e., an operational difference-maker to help IBM become adaptive and responsive, gain market share, reduce costs, grow revenue and profit, improve cash flow, and enhance client satisfaction.

Today, IBM's supply chain is managed on a global basis, leveraging costs through an integrated network of global suppliers and partners. The ISC encompasses manufacturing, procurement, customer fulfillment, and global logistics and includes nearly 20,000 employees spread across 56 countries. In 2006, the head of procurement, a major element within ISC, was relocated to Shenzhen, China, from the corporate headquarters in Armonk, New York.

Within ISC, the customer fulfillment process offers a good example of the benefits of global integration. Its transformation began in the early 1990s, just as the Internet was transforming the way individuals and organizations work. IBM began to extend electronic links for collaboration to suppliers, partners, and clients to streamline its process, improve its visibility into the supply chains, reduce inventory, and enhance its collaboration with their partners. The new single globally integrated operation immediately started realizing savings of 10–15% year after year. ISC eliminated steps in some fulfillment processes and automated others: For example, the order process was streamlined, eliminating the redundancy of having clients order from business partners and then business partners order from IBM. Today, clients order directly from business partners' Web sites using B2B systems with orders automatically feeding into the IBM order system. In fact, 95% of orders through business partners in the United States are automated. Client support processes have

been automated through a variety of Web tools, telewebs, and self-service applications that enhance client satisfaction, reduce support cost, and improve productivity. For example, the combined effort cut the average processing time for a purchase order from a month to a few hours, driving substantial savings in the form of paperless processes and automation.

Through this transformation and the use of business analytics, IBM componentized the customer fulfillment processes, deciding which process steps were best done close to the client and which ones could be handled globally. This assessment led IBM to extract transaction processing and data entry work and consolidate it in global delivery centers in Malaysia, Slovakia, Spain, and Brazil. As a result, roughly 20% of customer fulfillment resources are in low-cost countries. For the other 80%, these resources have been redirected toward higher value work, closer to client teams. For example, in Europe, customer fulfillment resources are working on high-value tasks and new roles such as customer relationship and proposal team coordinator. This has helped reduced the time that sales teams spend on fulfillment activities by 25%, allowing them to spend nearly 40% more time with clients.

As the globally integrated supply chain became a model of integration for IBM, the company began applying the experience to all its operations. For example, supply-chain principles and tools are being adapted to apply to managing hardware and software assets to increase competitiveness in the services business. In addition, IBM also takes its ISC know-how to help its clients manage and improve their own supply chains.

One of the key areas in which IBM differentiates itself and takes the lead in the industry is its use of Operations Research/Business Analytics (OR/BA) and information technology coupled with innovative business models and disciplined business process reengineering to transform its supply chain into a "smarter supply chain" to achieve a competitive advantage. The innovative use of OR along with process and information technology in the following four interrelated areas have supported IBM's supply-chain transformation in the last 15 years:

- *Extended Enterprise Supply-Chain Management*
- *Innovative Business Models and Business Optimization*
- *Adaptive Sense-and-Respond Value Net*
- *Value-Driven Sales and Delivery*

11.4.1 Extended Enterprise Supply-Chain Management

In 1993, IBM launched an internal reengineering effort to streamline business processes. The reengineering effort focused on improving customer satisfaction and market competitiveness by increasing the speed, reliability, and efficiency with which IBM delivers products to the marketplace. In 1994, the company added an asset management reengineering initiative to the effort.

A cross-functional team identified five areas that needed modeling support: (1) design of methods for reducing inventory within each business unit;

(2) development of alternatives for achieving inventory objectives for senior management consideration; (3) development and implementation of a consistent process for managing inventory and customer-service targets—including tool deployment—within each business unit; (4) complete evaluation of such assets as service parts, production materials, and finished goods in the global supply network; and (5) evaluation of cross-brand product and unit synergy to improve the management of inventory and risk. The Asset Management Tool (AMT), an OR-based strategic decision-support tool, was developed to address these issues. AMT integrates graphical process modeling, analytical performance optimization, simulation, activity-based costing, and enterprise database connectivity into a system that allows quantitative analysis of extended supply chains. The central function of the optimization engine is a constrained multi-echelon inventory optimization model for large-scale supply networks which couple nonlinear programming with gradient search, heuristic clustering, and queuing analysis [15]. IBM has used AMT to study such issues as inventory budgets, turnover objectives, customer-service targets, product simplification, and new-product introductions.

This work became the backbone of the successful reengineering of many IBM business units in North America and Europe, as well as for customers such as GE Capital, Best Buy, and Xilinx [38]. Financial savings through the AMT implementations amounted to more than $750 million at IBM Personal System Group in 1998 alone. Furthermore, AMT has helped IBM's business partners to meet their customers' requirements with much lower inventory and has led to a co-location policy with many business partners. In 1999, The IBM AMT team received the 1999 INFORMS Franz Edelman Award as well as IBM's Outstanding Technical Achievement Award.

11.4.2 Innovative Business Models and Business Optimization

In early 2000, a major IBM effort was to transition from an indirect build-to-plan business model to a flexible build-to-order and configure-to-order (CTO) business model to support a hybrid indirect and direct (Web-based) business [11]. A configure-to-order (CTO) system is a hybrid of build-to-plan and build-to-order operations. In a traditional build-to-plan or build-to-order environment, there usually is a pre-fixed set of end-product types from which customers must choose, as well as a pre-specified notion of demand types. In contrast, a CTO system allows each customer to configure his/her own product in terms of selecting a personalized set of components that go into the product. Therefore, the CTO system appears to be an ideal business process model that provides both mass customization and a quick response time to order fulfillment.

To support this transition, a set of OR-based initiatives were formed to perform supply-chain assessment, examine and enhance business processes, and optimize supply-chain policies and control parameters in the CTO environment [4]. Many interesting and new OR problems such as building-block-based forecasting [20],

[19], pricing, [1], [8] inventory optimization for CTO products [5], flexible supply contract, and reverse logistics [39] were discovered and analyzed. One example was about exploring flexible supply contracts as a means to facilitate coordination among supply partners [12]. Properly designed supply contracts allow value-net partners to share demand-and-supply risks and enable better coordination between decentralized supply chains while lowering costs. Quantity flexibility can be specified in a supply contract to allow a buyer to adjust its order quantities after the initial order is placed. Such flexibility enables buyers to reduce its risk of overstock or understock, which naturally comes at an extra cost to buyers. The extra cost gives the supplier incentive to offer flexibility while undertaking more risk. The model also generates qualitative insights to support channel coordination through a profit-sharing mechanism. This kind of analysis can be leveraged to evaluate the shared risks and fair compensations in a globally integrated supply network for GIEs.

Recognizing the value of OR/Business Analytics to business, the Value Chain Innovation Center (VCIC) was formed in 2002 with support from both IBM ISC and IBM Research. The mission of this center has been to create a cross-business and cross-functional "incubator" to develop advanced technologies and thought leadership for value-net collaboration and optimization, to create a value-net community, and to build a knowledge repository for assets. This center became the key technology center for delivering advanced technologies for ISC value-net transformation and is still actively supporting ISC technology needs today.

11.4.3 Adaptive Sense-and-Respond Value Net

Following the transition from Build-to-Plan to a hybrid of Build-to-Order and Configure-to-Order business model, it was time to explore more flexible and responsive models to help IBM leapfrog the competition. As discussed in Section 11.3 above, the Sense-and-Respond Value-Net effort was first initiated in 2000 to build an open and adaptive framework, using intelligent decision making and IT technology for business optimization. There have been several successful S&R pilots/implementations since then. Sense-and-Respond Value Net was adopted by IBM as a key supply-chain strategy in 2003. We will discuss two S&R implementations below: Sense-and-Respond Demand Conditioning and Virtual Command Center.

11.4.4 Sense-and-Respond Demand Conditioning

Sense-and-Respond Demand/Supply Conditioning Solution enables the supply chain to sense fluctuations in demand early on, intelligently analyze the signals, and seamlessly adjust itself in real time [6]. It allows a better understanding of transactional data representing customer needs, provides visibility of real-time supply-and-demand conditions, identifies supply–demand imbalances, and indicates out-of-threshold situations on an enterprise dashboard to allow proactive decision

making and needed adjustments. The system analyzes the order loads, shipments, supply commits, and demand forecasts data from enterprise-planning systems, correlates and analyzes the information, identifies imbalance events, alerts the appropriate business users, and recommends corrective actions. A key analytics system is the Order-Trend Analysis. "Order Analyzer" uses both historical demand and partial demand signals that are visible in a current time period, as well as other demand-related signals that can serve as headlights for future demand.

The implementation of S&R Demand Conditioning at IBM Personal Computing Division (PCD) in 2004 has produced great business benefits and improvements, including better data integration and visibility for earlier, more efficient responses and fast resolution. The order-trend analysis gives PCD earlier headlights into customer needs and supply constraints and excesses. Before the new process was implemented, demand-and-supply imbalance would need to contact each function separately, identify solutions to imbalances, and reach consensus on the best solution. The resolution of an imbalance issue that could take as long as 2 months was dramatically reduced. Sales also became more efficient and overall sales volume has increased through improved product availability or substitution. In the last quarter of 2004, time spent on administrative activities declined by 20% and sales increased by 5%. In addition, there was a 40% reduction in unfilled orders worldwide with $200 million in additional revenue.

Virtual Command Center

The Virtual Command Center (VCC) is a multi-enterprise, supply–demand balancing and collaboration solution based on the S&R model. It is composed of three major hubs which manage and synchronize demand, supply, and logistics (see Figure 11.4). It offers visibility, real-time performance management, event management, collaboration enablement, analytical platform, and intelligence. IBM is currently using the VCC Demand Hub in its own supply chains so as to collaborate with channel partners in order to support smart alignment of demand and inventory supply decisions and execution for selected products in North America and Europe [26].

Three key analytical capabilities were developed and incorporated. The Channel Sales Forecasting function predicts demand at business partners, analyzes entire sales out profile, incorporates headlights such as future marketing campaign data and promotion, detects abnormal events that deviate significantly from historical profile, and captures order skew by placing larger weights on historical sales in the same week within a quarter. The Optimized Buy Recommendations function captures price protection expenses, inventory carrying costs, and customer service; analyzes "lumpiness" of historical sales out; and minimizes costs while achieving a target service level (98% product availability at distributor). Finally, the Demand Shaping function identifies viable product alternatives if preferred product choices are unavailable to support "sell-what-you-have."

The business benefits of VCC have been significant. Within 1 year of VCC implementation along with related business process transformation initiatives, the total inventory is now down by 50% for the selected products in the United States. Promotion payments and price protection payouts were also reduced. The VCC has been deployed in more than 20 countries with more than 40 distributors in North America and Europe with more and more business partners increasingly accepting the VCC's purchase recommendations.

11.4.5 Value-Driven Services and Delivery

In recent years, IBM has undergone a transformation from a hardware company into a major services and software business. The company's revenue from Services has increased from $11B in 1993 to $59B in 2008. In 2005, a key effort, Value-Driven Sales and Delivery (VDSD)—later renamed the "Financial Transformation Workbench" (FTW)—was initiated to support Service Sales and Delivery [32]. The motivation was that enterprises increasingly focus on delivered values rather than on product, function, or initiatives. When buying an external service, enterprises expect the service provider to demonstrate the value of its services throughout the sale and delivery phases. If it is an internal initiative, they expect to see value before, during, and after implementation. Figure 11.5 shows the VDSD model. It provides an environment for enterprise-wide capability assessment and a comprehensive framework for design, development, deployment, and operation of services/initiatives.

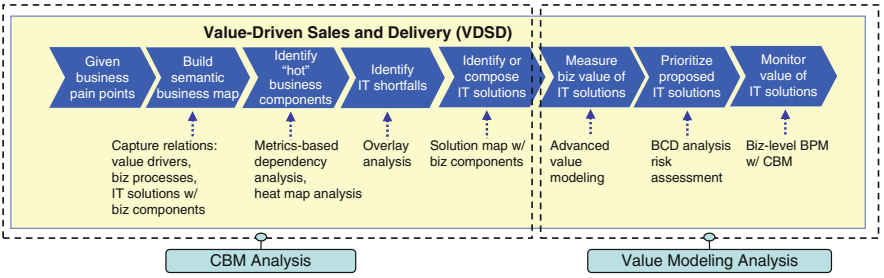


Fig. 11.5 Value-driven services and delivery model (Source: [32])

This model combines OR quantitative modeling with component-based qualitative modeling to help enhance sales and services based on business value. More specifically, VDSD leverages advanced business modeling techniques including Component Business Modeling (CBM) and semantic modeling and value modeling to assist customers in identifying areas for business transformation and operational improvement, recognizing and categorizing deficiencies in existing IT systems, calculating business values of transformational and IT initiatives, and prioritizing IT initiatives based on business values.

At the forefront of the industry trend of focusing on value in sales and services [25], VDSD is a pioneering effort of integrating business, IT and delivery, risk analysis and management, and of creating tools that link, calculate, monitor, and demonstrate value delivered. An IBM research and service team filed five VDSD patents [32]. CNN reported that "This Finance Transformation Workbench tool underscores the future of IBM's services business. The services' model of the future includes analytical software coupled with high-value consulting services and world-class research underpinning it" [CNN News, July 2008].

IBM has demonstrated a core competence in business transformation. It has created a new business model—the GIE—and delivered significant financial performance. State-of-the-art business analytics and information technology have been used through the transformation journey to help enable growth and productivity. The resulting transformation showcases how science was brought to the art of decision making to help optimize business performance. However, culture change always plays a key role in any business transformation.

11.5 Sustainability of the Globally Integrated Enterprise

Sustainable GIEs are enterprises that participate in global commerce and leverage global resources and capabilities to improve their business performance smartly while reconciling their needs with those of their supporting ecosystems from an environmental, social, and economic perspective. They are often *Globally Distributed and Economy Driven*; *Integrated*; *Agile*; *Performance Driven and Technology Enabled*; *Skills*, *Innovation, and People Focused*; and *Environmentally, Socially, and Economically Responsible*.

1. *Globally Distributed and Economy Driven* Competition has forced companies to seek global markets and operations but globalization has also greatly increased complexity and risks. With ever-increasing competition and narrowing operating margins, it has become more important than ever for companies to understand end-to-end performance, to make intelligent use of available resources, and to invest in moving their operations to where they will be most cost effective. It may also require decomposing the company into modular functions according to needs. These modular functions can be either supported within the organization or outsourced to different areas.

   A GIE therefore needs to strategically distribute its operations globally and the distributed entities need to perform their operations efficiently as an integrated enterprise. This requires the GIE to know its own capabilities and those of its partners as well as the values and costs/risks of each potential participant or solution component. It needs to evaluate the merits/impacts of the operational design and the adjustments needed to improve business performance. The Value-Driven Sales and Services-Delivery model (VDSD) that we describe in Section 11.4 can be used to model the values and costs of value-net participants and help an enterprise design its GIE operational model.

2. *Integrated* A key challenge that GIEs face is to integrate distributed operations and get partners across the globe to work in tandem despite the difficulties introduced by time, distance, communication, and culture barriers. To operate efficiently, participants in the value net need to share critical information to improve the situational awareness of the entire value net as well as that of the environment in which they operate. They also need to coordinate and synchronize their operations and collaboratively make decisions to address unexpected events. In the event of a supply-chain disruption, real-time assessment of the impact across the value chain is crucial for corrective actions. The IBM Virtual Command Center is a supply-chain/value-net solution that not only visualizes but also manages supply-chain visibility and real-time events based on integrated information. It was designed for collecting and integrating information from a heterogeneous global environment of business units and value-net partners. Its three major hubs manage and synchronize demand, supply, and logistics needs and provide analytics that greatly enhance related real-time decision making for harmonizing these needs.
3. *Agile* In the new global, continuously changing environment, events such as financial market disruptions, customer buying behavior changes, pandemic threats, terrorist attacks, and natural disasters, once considered rare, are becoming more commonplace. Disruptive technologies are also increasingly affecting business. Companies can no longer ignore the threats of a changing environment and need to prepare themselves to effectively adopt new technology for evolving situations. In this fast-changing environment, agility is a critical capability for an enterprise to remain sustainable. The faster a company can change its operational model to adjust to environmental changes, the more competitive it becomes. However, enabling a large enterprise to become agile is no simple task. The business processes need to be streamlined and, more importantly, the operational model needs to be flexible for quick reconfiguration. The changes need to be automated with application support such as the performance-, model-, and value-driven VDSD described in detail in Section 11.4.
4. *Performance Driven and Technology Enabled* A GIE employs Communication, Operation Research, Business Analytics, and Information Technologies to improve its business performance and to react to environmental changes. In addition, business leaders are looking for technology that will help them analyze large amounts of data collected from different sources so that they can, proactively and, if possible, in real time, detect exceptions and conduct root-cause analysis quickly and effectively and make an optimal use of resources. They seek technology that will generate alerts and quickly communicate those alerts to concerned parties [32]. GIEs need technology support for accurate and timely performance reports, disruptive events recognition, and role-based event notification with integrated information that can be presented to executives, managers, and operators ensuring their timely and fast communication and action [33]. The VDSD model-driven framework enables rapid process and application integration at build time and performance monitoring and quick operation reconfiguration at runtime.

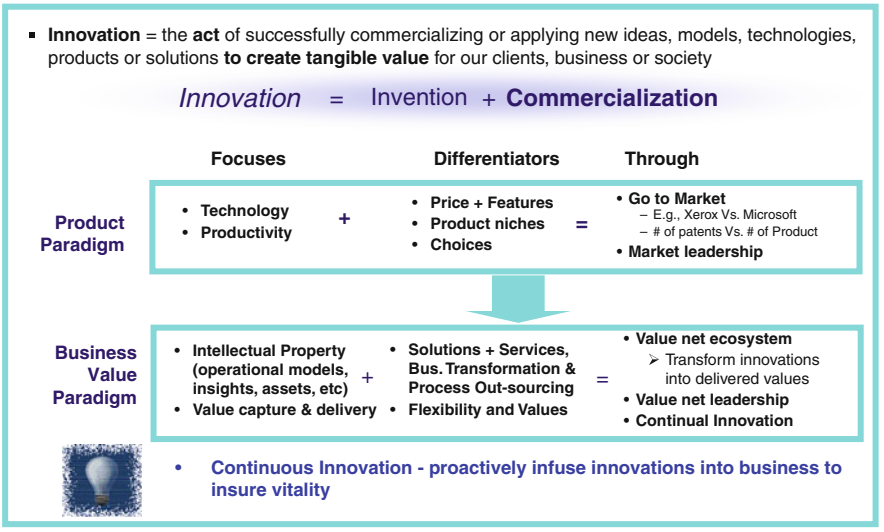


Fig. 11.6 Continual innovation is a fundamental source of competitive advantage (Source: [52])

5. *Skills, Innovation, and People Focused* An efficient GIE will continue to train its workforce and develop a culture of continuous innovation so it can remain the leader in its field (see Figure 11.6). It also has to pay attention to how globalization and the e-commerce transformation of the last 20 years have changed the enterprise landscape. Companies are increasingly forming value nets to collaborate with partners and clients, gaining shared situation awareness, and making quick decisions. This has created a need for state-of-the-art negotiation models and frameworks that can facilitate collaboration among partners.

   A company's most important resources are its people, skills, and assets. History shows that a market leader can fail quickly if it stops innovating or fails to sense and respond to market changes. IBM, Xerox, Kodak, Polaroid, Sears, Timex, US Steel, etc., are just a few of many great companies which once dominated their industries but then fell from grace because they stayed with their successful model for too long and failed to change with the environment. Some of them became great again by reinventing themselves but others faded into history. For an enterprise to be sustainable, it needs to reinvent itself continuously. In [17], the authors pointed out that the reason Dell's supply-chain success was difficult to replicate elsewhere was the company's culture and people. A culture of innovation can only take root when the company focuses on its people and encourages employees and partners to take risks to explore business innovation and sustainability. An innovative culture encourages taking calculated risks—even when these may result in occasional failure. Furthermore, the corporation relies on its people in all countries to self-regulate its operations, to be socially responsible, and to have a positive impact on the environment, communities, consumers, stakeholders, and employees.

6. *Environmentally, Socially, and Economically Responsible* For a company to transform to a GIE is a complex issue. Cost and skills are not necessarily the only considerations; environmental, social, and economic impacts are also critical success factors. Shifting operations can be costly and time consuming and issues and concerns need to be thoroughly analyzed and confronted. These issues range from transportation and distribution costs of physical goods and parts and the potentially positive or negative productivity impact of reduced collaboration caused by time-zone differences and distances to the much larger and sensitive geo-social-political issues of outsourcing jobs.

   The benefits of the GIE transformation are not always obvious and can be negated by the adverse impact of the transformation—particularly as related to economic, societal, and environmental factors. For example, shifting jobs to lower cost countries often reduces domestic jobs opportunities and brings about social/political issues which may affect domestic buying power or customer relationships, thereby lowering the demand for goods. On the other hand, the exploitation of low-cost skill pools often improves the local economy thereby increasing local buying power but, at the same time, increasing labor and other costs. The six characteristics of a well-run GIE that we have just discussed also apply to the sustainability of the GIE.

Corporate social responsibility is not a new concept but in the past companies have tended to focus on financial performance and only recently realized that short-term financial gain at the expense of product safety, social, and environmental responsibilities can have a long-term negative impact on their brand and business. The 2010 Toyota gas pedals and brakes issues, the 2007 Mattel toxic toys incident, the industrial accident of Union Carbide at the Bhopal, India plant, and the Exxon's Valdez oil spill in Alaska in 1989 all caused significant damage to these companies' business and brand equity.

More and more, companies are realizing that they can earn a profit while being socially and environmentally responsible. For example, British retailer Marks & Spencer (M&S) has embarked on a £200-million, 5-year plan that impacts almost every aspect of its operations. One initiative is to simultaneously improve efficiency and sustainability through its online supplier exchange. For instance, farmers who create biogases from farm waste are now selling green electricity to M&S—along with their beef. M&S has proven that it is possible to do well while doing good: the company's operating profit has increased at a compound annual growth rate of more than 14% for 5 consecutive years [42]. Starbucks and many other corporations' support of fair-trade coffee and tea have helped both the farmers and the company's business. Carbon trading is another example. A recent study found that there is a correlation between social/environmental performance and financial performance [44]. IBM's Smarter Planet initiatives have identified many areas where the companies can reap financial gains while being socially and environmentally responsible [Smarter Planet, Web]. Smarter city, Smarter Grid/Meters, Smarter Supply Chains, Smart Water Management, Smart Health Care, Green Data Centers, etc., are just a few of the promising examples.

11.6 Conclusion

We live in a hugely complex and interconnected world where the old criteria for maintaining a thriving and profitable business no longer apply. Taking the road to transformation creates immeasurable challenges, with great creativity and innovation as a result. Adapting to evolving technology and to different environments has created highly efficient new models. The new emerging model is the Globally Integrated Enterprise (GIE) that shows the way for companies to run and operate their global supply-and-demand chains. Our own research and the IBM experience in becoming a Sense-and-Respond GIE demonstrate how a deep awareness on the part of businesses to go beyond the bottom line and become responsible players on the global scene is helping companies take the road to transformation. Challenges remain but opportunities abound. Using the available and continuously refined Operations Research, business analytics, value-driven methods and tools, and information technologies, GIEs can become more socially, economically, and environmentally responsible and achieve sustainable success.

References

1. Bichler M, Kalagnanam J, Katircioglu K, King A, Lawrence R, Lee HS, Lin G, Lu Y (2002) Application of flexible pricing in B-to-B electronic commerce. IBM System Journal 41(2)
2. Bittner E (2000) E-business requires supply chain event management. AMR Research Report
3. Bradley S, Richard P, Nolan L (eds) (1998) Sense and respond: Capturing value in the network era. HBS Press Book, Massachusetts
4. Breitwieser R, Cheng F, Eagen J, Ettl M, Lin G (2000) Product hardware complexity and its impact on inventory and customer on-time delivery. International Journal of Flexible Manufacturing Systems 12(2–3):145–163
5. Brown A, Ettl M, Lin G, Petrakian R, Yao D (2001) Inventory allocation at a semiconductor company: modeling and optimization. In: Song JS, Yao DD (eds) Supply chain structures: Coordination, information, and optimization. Kluwer, Massachusetts
6. Buckley S, Ettl M, Lin G, Wang K (2005) Intelligent business performance management—sense and respond value net optimization. In: Fromm H, An C (eds) Supply chain management on demand. pp. 309–337, Springer, Massachusetts
7. Buckley S, Ettl M, Lin G, Wang KY (2005) Sense and respond business performance management. In: An C, Fromm H (eds) Supply chain management on demand, pp. 287–311, Springer
8. Cao H, Gung R, Lin G, Jang Y, Lawrence R (2006) Method and structure for bid winning probability estimation and pricing model. US patent 7139733. Issued December 2006
9. Castano-Pardo A, Lin G, Williams T (2006) On the move—Advancing military logistics toward sense-and-respond. IBM Industry Business Value Whitepaper
10. Cecere L, Hofman D, Martin R, Preslan L (2005) The handbook for becoming demand driven. AMR Research
11. Cheng F, Ettl M, Lin G, Yao D (2002) Inventory-service optimization in configure-to-order systems. Journal of Manufacturing Service Operations Management
12. Cheng F, Ettl M, Lin G, Schwarz M, Yao D (2009) Flexible supply contracts via options. In: Uzsoy R et al (eds) Handbook of production planning. Springer, Massachusetts
13. Correia J, Schroder N, Bam N (2002) A composite market view. Gartner report

14. Donofrio N (2009) Creating real value through innovation, transformation and continuous change. Asean CIO leadership exchange. July 2009. http://ftp.software.ibm.com/software/sg/cioleadershipexchange/0910hrs_NDD_ASEAN_CIO_Presentation_Final.pdf
15. Ettl M, Feigin G, Lin G, Yao D (2000) A supply network model with base-stock control and service requirements. Operations Research 48(2):216–232
16. Friscia T, O'Marah K, Hofman D (2007) The AMR research supply chain top 25 for 2007. AMR Research, May 31, 2007
17. Fugate B, Metzer J (2004) Dell's supply chain DNA. Supply Chain Management Review, October 1, 2004
18. Gerstner L (2002) Who says elephants can't dance? Inside IBM's historic turnaround. Collins, ISBN-10: 0060523794
19. Gung R, Hosking J, Lin G, Tajima A (2004) Demand planning for configure-to-order and building blocks-based market environment. US patent 6816839. Issued November 11, 2004
20. Gung R, Leung Y, Lin G, Tsai R (2002) Demand forecasting today. ORMS Today
21. Haeckel S (1999) Adaptive enterprise: Creating and leading sense-and-respond organizations. Harvard Business School Press, Cambridge
22. IBM CEO Study (2008) 2008 CEO study report—Enterprise of the future. May 25, 2008. http://www.ibm.com/ibm/ideasfromibm/us/ceo/20080505/
23. IBM Smarter Planet (20090 Smarter planet. Nov. 1, 2009. http://www.ibm.com/smarterplanet/
24. IBM (2005) Business impact of outsourcing—a fact based analysis. IBM Research, Research report. www.ibm.com/services/in/igs/pdf/ibm_biz_impact_of_outsourcing.pdf.
25. Kaplan R, Norton D (1992) The balanced scorecard—measures that drive performance. Harvard Business Review, Cambridge
26. Kapoor S, Binney B, Buckley S, Chang H, Chao T, Ettl M, Luddy EN, Ravi RK, Yang J (2007) Sense-and-respond supply chain using model-driven techniques. IBM Systems Journal 46(4)
27. Kumaran S (2004) Model driven enterprise. Global EAI summit, Banff, Canada
28. Lawrie G (2003) Preparing for adaptive supply networks. Forrester Report
29. Lee H (2004) The triple—A supply chain. Harvard Business Review 82(10):102–112
30. Lee H, Continuous A (2002) Sustainable improvement through supply chain performance management. Stanford global supply chain forum, Stanford university
31. Lee H, Peleg B, Whang S (2005) Toyota: Demand chain management. Mar 18, 2005. GS42-PDF-ENG, Stanford supply chain forum, Stanford university
32. Lee J, Lin G, Wang K, Woody C (2005) Value-driven sales and delivery. IBM white paper
33. Lee J, Lin G, Jang Y, Yao D (2005) System and method for value evaluation of business and it capabilities. Disclosure number: END820050173, Docket number: END920050106US1. Filed: August 2005
34. Lee J, Lin G, Wang K, Woody C (2005) System and methods for value-driven sales and delivery. Doc number: END9-2005-0068. Filed: August 2005
35. Lehmann C (2003) The rapid sense-and-respond enterprise: Part 1 & part 2. Meta group reports
36. Lin G, Luby B, Wang K (2004) New model for military transformation. OR/MS Today 31(6)
37. Lin G, Jeng JJ, Wang K (2004) Enabling value net collaborations. In: Chang YS et al. (eds) Evaluation of supply chain management. Kluwer, pp. 417–430
38. Lin G, Ettl M, Buckley S, Bagchi S, Yao D, Naccarato BL, Allan R, Kim K, Koenig L (2000) Extended-enterprise supply-chain management at ibm personal systems group and other divisions. Interfaces 30(1):7–21
39. Lin G, Buckley S, Cao H, Caswell N, Ettl M, Kapoor S, Koenig L, Katircioglu K, Nigam A, Ramachandran B, Wang K (2002) The sense and respond enterprise. OR/MS Today 29(2): 34–39
40. Lin G, Lu Y, Yao D (2008) The stochastic knapsack revisited: Switch over policies and dynamic pricing. Operations Research 56(4):945–957
41. Lin G, Wang K (2008) Presentation to the IBM Automotive Industry Architect Community. July 27

42. Marks & Spencer (2006 and 2007) Annual reports
43. Neubarth M (2009) The mainframe: The dinosaur that wouldn't die. Sept. 24, 2009. http://www.ciozone.com/index.php/Server-Technology-Zone/The-Mainframe-The-Dinosaur-That-Wouldn-t-Die/1.html
44. Orlitzky M, Schmidt FL, Rynes SL (2003) Corporate social and financial performance: a meta-analysis. Organization Studies 24(3):403–441
45. Palmisano S (2006) The globally integrated enterprise. Foreign Affairs Magazine 85(3) (May/June):127–136
46. Palmisano S (2008) A smarter planet—the next leadership agenda. Speech at the Council on Foreign Relations, New York. Nov 06, 2008.http://www.ibm.com/ibm/ideasfromibm/us/smartplanet/20081106/sjp_speech.shtml
47. Plambeck E, Denend L (2007) Wal-Mart's sustainability strategy. Case no. OIT71, Stanford supply chain forum, Stanford university.
48. Richards C (2008) The "Smarts" of an Intelligent grid: Analytics for intelligent grid initiatives. IDC Report #EQ12029, ISBN-10B001BXWEHM
49. Roehrig P (2007) Outsourcing clients can expect 12 to 17% savings. Forrester Research, Inc., August 30, 2007
50. SmartCity (2009) SmartCity. Dec. 10, 2009. http://www.smartcity.ae/
51. Suleski J, Quirk C (2001) Supply chain event management: The antidote for next year's supply chain pain. AMR Research Report
52. Wang K (2004) Creating a Sense and Respond Enterprise of the Future, Today. IBM 2004 Professional Technical Leadership Exchange

Chapter 12
Cyberinfrastructure and Optimization

Robert Fourer

Abstract In 2002 the U.S. National Science Foundation created a Blue-Ribbon Advisory Panel on Cyberinfrastructure, which submitted in January of 2003 a report entitled "Revolutionizing Science and Engineering Through Cyberinfrastructure." Subsequently, the NSF created an Office of Cyberinfrastructure (OCI) independent of its directorates in such traditional areas as biology, computer science, geosciences, physical science, and engineering. In the following 3 years the NSF sponsored workshops leading to nearly 30 reports (www.nsf.gov/od/oci/reports.jsp) on the role of cyberinfrastructure in specific areas of research. This chapter describes a variety of projects that fall into the intersection of cyberinfrastructure with the study and practice of large-scale optimization. In general, these projects involve large-scale optimization problems in system design, production planning, and logistics. However, the notion of large-scale optimization occurs in other disciplines including physical and biological sciences, engineering, economics. As such, there is a benefit to establish a community whose members use the same modeling and algorithmic techniques and who can benefit from the same software and services.

In 2002 the U.S. National Science Foundation created a Blue-Ribbon Advisory Panel on Cyberinfrastructure, which submitted in January of 2003 a report entitled "Revolutionizing Science and Engineering Through Cyberinfrastructure" [2]. Subsequently, the NSF created an Office of Cyberinfrastructure (OCI) independent of its directorates in such traditional areas as biology, computer science, geosciences, physical science, and engineering. In the following 3 years the NSF sponsored workshops leading to nearly 30 reports (www.nsf.gov/od/oci/reports.jsp) on the role of cyberinfrastructure in specific areas of research.

OCI's statements of its mission (www.nsf.gov/od/oci/about.jsp) provide a taste of what the term *cyberinfrastructure* is intended to encompass:

Robert Fourer
Northwestern University, Evanston, IL 60208, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_12,

> The Office of Cyberinfrastructure coordinates and supports the acquisition, development and provision of state-of-the-art cyberinfrastructure resources, tools and services essential to the conduct of 21st century science and engineering research and education.
>
> OCI supports cyberinfrastructure resources, tools and related services such as supercomputers, high-capacity mass-storage systems, system software suites and programming environments, scalable interactive visualization tools, productivity software libraries and tools, large-scale data repositories and digitized scientific data management systems, networks of various reach and granularity and an array of software tools and services that hide the complexities and heterogeneity of contemporary cyberinfrastructure while seeking to provide ubiquitous access and enhanced usability.
>
> OCI supports the preparation and training of current and future generations of researchers and educators to use cyberinfrastructure to further their research and education goals, while also supporting the scientific and engineering professionals who create and maintain these IT-based resources and systems and who provide essential customer services to the national science and engineering user community.

The purpose of this chapter is to describe a variety of projects that fall into the intersection of cyberinfrastructure with the study and practice of large-scale optimization, as explained further in Section 12.1.

Of particular interest in this context are frameworks for making optimization software more readily available; Sections 12.2, 12.3, and 12.4 present distinct projects for this purpose. Several other projects, considered in Section 12.5, are related by the goal of helping people make better use of available optimization software. Finally, Section 12.6 describes efforts to apply diverse high-performance computing facilities to problems of optimization. Concluding remarks in Section 12.7 see these activities as having an encouraging future, though perhaps less as the sort of cyberinfrastructure projects that appeal to research sponsors (such as NSF's OCI) and more in the context of emerging business models that are beginning to show promise.

Naturally many of these projects have to do with minimizing costs or maximizing profits in operations research applications. A great variety of activities in design, manufacturing, distribution, and scheduling seek to minimize costs or maximize profits (or surrogates for these). But optimization is also an established paradigm for problems in the physical and biological sciences, numerous engineering disciplines, economics, and business, ranging as broadly as the minimization of energy in a protein structure, the cost of a circuit configuration, and the total bid price of a combinatorial auction. Problems of these and many other kinds are addressed by a large optimization community whose members use the same modeling and algorithmic techniques and who can benefit from the same software and services.

12.1 Cyberinfrastructure and Optimization

Everyone is familiar with infrastructures: road systems, rail networks, power grids. An infrastructure does not produce goods or services itself; rather, it makes a wide range of productive activities possible. The interstate highway infrastructure does

not itself carry out supply-chain management, for example, but it permits the development of supply-chain management systems that would not be possible otherwise. Indeed, it paves the way for phenomena that were not foreseen when it was built, such as crossdocks and suburban sprawl. The effectiveness of infrastructures depends critically on standards (track gauges and time zones for railroads, bridge heights for highways, voltages for power grids) and on accessibility to a broad base of users.

Among the major infrastructures of modern life, cyberinfrastructures constructed from computers, data networks, software, and communication standards are among the newest and most elaborate instances. The Internet and the Web are the best known examples. Like other infrastructures, they facilitate myriad applications—the Web's use for unexpected purposes is already legendary—and they depend critically on software standards such as IP, HTTP, and HTML.

Optimization as currently practiced is inherently computational. Of greatest relevance to cyberinfrastructure are optimization software packages that address problem classes defined by mathematical properties of the objective and constraints, such as linearity or discreteness of the variables. Hundreds of these *solvers* are in regular use, based on a broad variety of optimizing algorithms and combinations of algorithms, many of them quite complex; each offers some trade-off between breadth of problem, efficiency of solution, convenience of implementation, and cost. At the same time a variety of *modeling languages* and support systems have been developed to translate between the problem representations familiar and convenient to human modelers and the data structures required for efficiency of the algorithms. The independent profusion of solvers and of modeling systems is characteristic of optimization and provides much of the impetus for the creation of independent optimization infrastructures.

Indeed, solvers resemble infrastructure tools in several respects. They do not directly address people's concerns in science, engineering, or commerce, but rather serve as tools for bringing optimization models to bear within application areas and systems. As a result the concept of optimization has been applied to many problems that were unknown to the creators of the relevant solvers. At the same time optimization software has become more accessible through the adoption of interfaces that, although serving as standards only for certain problem types or product groups, are at least widely known and readily grasped by individuals who have technical training in many different fields. These characteristics, together with distribution through the Internet, underlie the possibilities for optimization cyberinfrastructures of diverse kinds.

Application-specific optimization software targets models and methods in particular areas of endeavor such as vehicle routing, pattern cutting, workforce scheduling, circuit design, or portfolio management (to name just a few). These kinds of optimization packages are used in relatively predictable ways and tend to be designed as self-contained "solutions" that have less need for standard interfaces. Nevertheless, these packages often use general-purpose solvers as components and in doing so can also make good use of optimization cyberinfrastructures.

12.2 COIN-OR

The COIN-OR Foundation (www.coin-or.org) manages an initiative to support the development of open-source software for the operations research community. Founded in 2000 as the *Common Optimization Interface for Operations Research,* its scope has broadened [14] and its name has been changed to the *Computational Infrastructure for Operations Research.* Nevertheless after 9 years its 35 projects are still predominantly in the optimization field.

COIN-OR acts as a cyberinfrastructure in several ways. It is an Internet repository for freely available, general-purpose solvers that can serve as foundations for optimization applications as previously described. It makes available uniform tools for developing, managing, and documenting open-source optimization projects. Indeed it provides tools at a number of levels, in a way that encourages building new solvers upon routines already available, both for specialized functions (automatic differentiation, cut generation) and for easier problems (linear programming). This focus distinguishes COIN-OR from larger, more general open-source repositories such as SourceForge or the GNU Free Software Directory.

As presented on the COIN-OR home page, open source has a number of attractive features as a paradigm for software development:

When people can read, redistribute, and modify the source code, software evolves. People improve it, people adapt it, people fix bugs. The results of open-source development have been remarkable. Community-based efforts to develop software under open-source licenses have produced high-quality, high-performance code—code on which much of the Internet is run.

This is an appealing context for the development of optimization software. Many COIN-OR solvers were initially developed in the context of research and then gradually improved through the addition or combination of algorithmic ideas. Additionally the IBM Corporation contributed many of the initial projects, which were considered worth the effort of development but perhaps too specialized to justify commercialization.

By requiring its projects to select from licenses approved by the Open Source Initiative (www.opensource.org), COIN-OR adopts an expansive view of open source that does not allow, for example, software that is only free for academic use. Open-source licenses do differ in the requirements that they impose on reuse and redistribution, and here COIN-OR has encouraged (though not required) the use of licenses that permit incorporation of its software into proprietary, non-open projects. This reflects COIN-OR's IBM origins as well as a general desire in the OR community to promote optimization methods as relevant to operational problems faced by industry.

12.3 The NEOS Server

Since 1996 a large group of collaborators have developed NEOS, a *Network-Enabled Optimization System*, with the goal of making optimization an Internet resource [4]. The NEOS Server (neos.mcs.anl.gov) in particular has become a key

online resource in the optimization field, not by providing solvers for download like COIN-OR, but rather by offering a *software service* that accepts descriptions of optimization problem instances and sends back solutions. A central server, established at Argonne National Laboratory, manages solver requests generated through specialized Web forms and submission tools; it maintains "job" queues, monitors progress, and returns results, while hosting guides to solver features and submission procedures. The work of running solvers is farmed out to other computers contributed at a variety of locations, so that the service is readily scaled up.

The NEOS Server has had a continuing impact on optimization research, teaching, and applications, by providing immediate access to over 60 solvers—far more than optimization users could hope to install locally. Many are open source (from COIN-OR and elsewhere), with a strong representation of algorithms based on recent research in such areas as global optimization, semidefinite programming, and nonlinear optimization over integer variables. But even commercial solver developers have made their products available free through NEOS to encourage potential customers to try them out. In 2009 submissions were averaging about 20,000 a month, predominantly using commercial modeling languages also provided free by their developers.

For the optimization community, the NEOS Server provides the characteristics generally associated with a cyberinfrastructure: facilitating applications rather than directly performing them; enabling more applications than were originally imagined; providing open access to Internet-based resources; and supporting whatever standards solvers have adopted for the expression of problems. Originally a standalone tool, the Server has adopted the eXtensible Markup Language (XML) standard for data transfer and XML/RPC [5] for remote procedure calls, so that its facilities can be invoked from programs running anywhere on the Internet.

The NEOS Server's success has largely been as a tool for learning, experimentation, and benchmarking. While there are no rules against its ongoing use in support of a project or business, it does not provide the guarantees of reliability or confidentiality that would encourage such applications. Its emphasis has reflected in part its origins, as in the case of COIN-OR, but those origins have been quite different: a team drawn from the numerical analysis community more than operations research and a focus at a government laboratory rather than a corporation (though academics were involved in both cases).

12.4 Optimization Services

Looking forward to a "next generation" of the NEOS Server, a newer project has been undertaken to design a distributed optimization environment in which modeling languages, servers, registries, agents, interfaces, analyzers, solvers, and simulation engines can be implemented as services and utilities under a unified framework. This work, called Optimization Services or OS (www.optimizationservices.org), defines standards for all activities necessary to support decentralized optimization on the Internet. A reference implementation [6] is freely available as an open-source project under COIN-OR.

The OS framework conceives of optimization as a modern software service, based on Internet-wide standards such as Web Services, Service-Oriented Architecture, and XML. Thus it is a specialized cyberinfrastructure, but unlike more traditional optimization systems it is designed explicitly to integrate optimization into broader distributed computing environments, using technologies that are already familiar to the Information Technology community.

Making optimization into the kind of service envisioned by the OS project is easier said than done. For one thing, optimization currently relies on a hodgepodge of not-quite-standard formats for problem description, some tracing their origins as far back as punch card technology. These formats are moreover entirely inadequate to the needs of powerful modeling languages and analysis tools; as a result each optimization modeling product has adopted its own proprietary scheme for representing problems and results. Whereas the NEOS Server leaves it to each solver host to decide what input formats to accept, the OS project incorporates an initiative to create a comprehensive standard, the OS instance language or OSiL, for representing linear, nonlinear, stochastic, and other broadly applicable problem instances in a consistent way. To meet the needs of varied optimization environments, OSiL specifies an XML-based file format, a corresponding in-memory data structure, and a common interface to these forms for data transfer and function evaluation.

The requirements of a comprehensive optimization service demand a variety of other standards and protocols, moreover, which scarcely exist at present. These serve purposes such as

- representation of solver algorithms' options and results;
- communication between clients and solvers;
- registration and discovery of solvers and related software using the concept of Web Services.

Designing such standards is particularly challenging because optimization services exhibit a greater variety and complexity of information to be moved around than do typical business applications. To further complicate matters, the mathematical problem types that categorize solvers do not readily correspond to the model types familiar to human users. Overall, building an OS framework is much more of a challenge than simply copying XML, SOA, and Web Services ideas from existing software over to optimization packages.

The OS project's ultimate goal is to "make optimization as easy as hooking up to the network." The vision is for all optimization system components to be implemented as services under the OS framework and for customers to use these computational services much like utilities, with specialized knowledge of optimization algorithms, problem types, and solver options being potentially valuable but not required. The OS framework will in turn be built upon standards that are independent of programming language, operating system, and hardware and that are open and readily available for use by the optimization community.

The OS project's success will necessarily depend on developers' acceptance of its proposed standards. COIN-OR's OSI project has shown one way of facilitating this standardization on the solver side, by creating a more uniform interface to linear

and mixed-integer solvers. But it will be a greater challenge to get products on the modeling language side to forego their proprietary interfaces, which have been tuned and specialized over the years, in favor of a standard representation of solver inputs and outputs. Initially such a change will only mean more work, but over the longer term it promises to streamline the creation and maintenance of solver interfaces.

12.5 Intelligent Optimization Systems

Optimization services have largely been conceived as providing solver access to people who seek optimal (or at least very good) solutions to optimization problem instances. Underlying this view has been a confidence that owners of problems are knowledgeable as to which solvers are appropriate. Yet as previously noted, solvers are applicable to specific mathematical problem types distinguished by technical characteristics such as linearity, smoothness, and various discrete and logical structures. These do not readily correspond to the concerns of modelers who are thinking in terms of production, distribution, scheduling, design, and other model types applied in particular areas of science, engineering, and commerce.

It is thus worth considering what might be gained by taking a broader view. One can imagine an optimization cyberinfrastructure that incorporates software to aid in the selection of solvers. Features might include converting common nondifferentiable and discontinuous functions to forms that diverse solvers can handle; identifying convexity, both generally in objective functions and constraint regions and specifically in the case of constraints that can be viewed as quadratic cones; and making natural combinatorial and logical operators accessible to both numerical and logic-based solvers. The DrAMPL project [9] has taken some steps along these lines, including the matching of deduced problem characteristics against a database of solver features.

Going further, one can envision a optimization services framework that incorporates "intelligent" assistance for modeling, tuning of solver options, and analysis of results. Software embodying aids for these purposes were in existence as far back as the late 1970s, when ANALYZE [12] was developed at the U.S. Federal Energy Administration. Greenberg [11] provides an overview and bibliography of developments through the mid-1990s.

Work in this area has continued, as evidenced by MProbe [3] which offers an extensive suite of analysis tools and graphics for examining the shape of the objective function, the effectiveness of constraints, and the characteristics of the feasible region. Other mechanisms for problem analysis and transformation are found increasingly in implementations of modeling languages and solvers. There remain many ways in which the power of such systems could be further expanded, however, and it will be a significant challenge even to adapt existing systems like MProbe to function as independent services that can be treated as part of the infrastructure of optimization.

12.6 Advanced Computing

Software as a service implies the existence of hardware platforms to act as servers. Current optimization service frameworks, like NEOS and OS, rely on ordinary computers, mainly PCs running Windows or Linux. But there also exists the potential to enhance the practice of optimization by bringing advanced computing—a concept widely associated with cyberinfrastructure—to the optimization community.

In the context of optimization, "advanced" may refer to any of several approaches that employ multiple processors to accomplish what cannot be done effectively by individual computers, including

- *high-performance computing,* exploiting large numbers of processors through specialized, high-speed interconnections;
- *distributed computing,* using conventional computers working together through standard networks;
- *high-throughput computing,* marshalling the computational resources of otherwise idle networked computers.

A great variety of optimization problems have features that permit advanced computing to be used to advantage. For example, the metaNEOS project of 1997–2001 applied advanced computing approaches in solving all of the following:

- the 10^{10}-variable deterministic equivalent of a 107-scenario stochastic program on a computational grid of about 800 workstations, in about 32 h of wall-clock time [13];
- a previously intractable quadratic assignment problem using an average of 650 worker machines over a 1-week period, providing the equivalent of almost 7 years of computation on a single workstation [1];
- a mixed-integer nonlinear programming problem with parallel efficiency approaching 80% on 600 million search-tree nodes [10].

Yet despite the impressive technical achievements of these and similar projects, they have had a disappointingly limited impact on optimization in practice. Indeed, experience in the use of advanced computing platforms remains rare among people trained in large-scale optimization. For most members of the optimization community, whose focus is modeling and solving rather than computing, it is a daunting challenge to arrange for the hardware and software resources necessary to apply or even experiment with such advanced computational approaches.

The software services concept offers a clear possibility for a remedy to this situation. An advanced computing platform and the software tailored to it could be set up to act as an optimization server. Users anywhere on the Internet could send their problems to be solved, in much the same forms as are sent to ordinary solvers through NEOS today, and requiring at most a limited knowledge of advanced computing technology. The developers and maintainers of the optimization methods implemented on such servers would need to understand the technology in detail, but they would see their efforts benefit a great many more applications than at present.

High-performance computers are already accessed by their users via the Internet, to be sure. But for reasons of scarcity, security, or simply custom, specialized multiprocessor computers and large multiprocessor networks have been available only by prearrangement of availability of the software and, in many cases, availability of hardware time. In contrast, optimization users expect to be able to request the use of algorithms when they are needed and for unpredictable amounts of time; that is the level of service available from NEOS, after all. Such needs are inherent in the nature of large-scale optimization, which involves the use of algorithms that work well in practice but have no theoretical performance guarantees and in fact exhibit performance that is highly variable (though quite good on average). For the hardest problems, variability is made even greater by the use of complex iterative schemes that repeatedly apply assorted algorithms to a range of automatically generated problems.

In sum, an infrastructure for large-scale optimization on advanced computing platforms will require a sort of *supercomputing on demand* that does not seem to have been so necessary for other applications. This is an area where the optimization and computing communities could benefit from collaboration on substantive and original cyberinfrastructure research.

12.7 Prospects for Cyberinfrastructure in Optimization

This chapter began by introducing its topic through a description of the National Science Foundation's Office of Cyberinfrastructure, whose mandate is to fund basic research. Do innovations in cyberinfrastructure for optimization have a potential to be treated as research contributions? Some of the work described herein has been funded by NSF and other agencies, though not directly by OCI. Yet grant panelists and journal referees have at times viewed these projects as straightforward applications of ideas already pioneered more broadly in the context of Information Technology. To advance cyberinfrastructure as a research topic in optimization, proponents of this area of investigation will have to better educate the IT and OR communities in the aspects of optimization that truly pose challenges for cyberinfrastructure projects. Some of these aspects have been noted in this chapter.

Perhaps the creation of cyberinfrastructures for optimization will evolve to be as much a commercial as a scientific activity, however. The last decade has seen an increasing number of companies that provide or embed optimization in their products and that could benefit from some of the ideas I have described. Bigger players such as SAS, Microsoft, and IBM are greatly expanding the role of optimization in their offerings and have the resources to establish the ideas and standards of the Optimization Services project among a broad range of clients.

Indeed many of the concepts described in this chapter have lately been brought together under the umbrella of "cloud computing," which is a predominantly commercial phenomenon. At least one large-scale solver is already being made available for a fee through Amazon's Elastic Compute Cloud facility (aws.amazon.com/ec2),

and this sort of development may further encourage efforts to bring solvers and modeling languages together as optimization services. Overall, the intersection of cyberinfrastructure and optimization would seem to have considerable potential for an exciting and influential future.

Acknowledgments

Several of the analyses herein have been adapted from earlier collaborative work on cyberinfrastructure [7, 8] and optimization services [6]. An earlier version of this chapter appeared in the Fall 2008 newsletter of the INFORMS Computing Society.

References

1. Anstreicher K, Brixius N, Goux J-P, Linderoth J (2002) Solving large quadratic assignment problems on computational grids. Mathematical Programming 91(3):563–588
2. Atkins DE, Droegemeier KK, Feldman SI, Garcia-Molina H, Klein ML, Messerschmitt DG, Messina P, Ostriker JP, Wright MH (2004) Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyber-infrastructure. Report cise051203. National Science Foundation. Available at www.nsf.gov/od/oci/reports/toc.jsp
3. Chinneck JW (2001) Analyzing mathematical programs using MProbe. Annals of Operations Research 104:33–48
4. Dolan ED, Fourer R, Moré JJ, Munson TS (2002) Optimization on the NEOS server. SIAM News 35(6):4, 8–9
5. Dolan ED, Fourer R, Goux J-P, Munson TS, Sarich J (2008) Kestrel: An interface from optimization modeling systems to the NEOS server. INFORMS Journal on Computing 20(4): 525–538
6. Fourer R, Ma J, Martin K (2008) Optimization services: A framework for distributed optimization. Technical report, Optimization Services Project. Available at www.er.org/CI/Optimization-Services.pdf; forthcoming in Operations Research
7. Fourer R, Moré JJ, Munson T, Leyffer S (2006) Extending a cyberinfrastructure to bring high-performance computing and advanced web services to the optimization community. Proposal to the National Science Foundation. Available at www.4er.org/CI/NSF-Proposal-2006.pdf
8. Fourer R, Moré JJ, Ramani K, Wright SJ (2004) An operations cyberinfrastructure: Using cyberinfrastructure and operations research to improve productivity in the enterprise. Report on a workshop sponsored by the National Science Foundation. Available at www.optimization-online.org/OCI/OCI.pdf
9. Fourer R, Orban D (2009) DrAmpl: A meta solver for optimization problem analysis. Computational Management Science, published online first, dx.doi.org/10.1007/s10287-009-0101-z
10. Goux J-P, Leyffer S (2002) Solving large MINLPs on computational grids. Optimization and Engineering 3(3):327–346
11. Greenberg HJ (1996) A bibliography for the development of an intelligent mathematical programming system. Annals of Operations Research 65:55–90

12. Greenberg HJ (1993) A computer-assisted analysis system for mathematical programming models and solutions: A user's guide for ANALYZE. Operations Research/Computer Science Interfaces Series, vol 1. Kluwer, Boston, MA
13. Linderoth JT, Wright SJ (2003) Implementing a decomposition algorithm for stochastic programming on a computational grid. Computational Optimization and Applications 24(2–3):207–250
14. Lougee-Heimer R (2008) COIN-OR in 2008. OR/MS Today 35(5):46. Available at www.lionhrtpub.com/orms/orms-10-08/frcoin-or.html

Chapter 13
Perspectives on Health-Care Resource Management Problems

Jonathan Turner, Sanjay Mehrotra, Mark S. Daskin

Abstract Research devoted to health-care applications has grown increasingly within operations research over the past 30 years, with over 200 presentations at the 2008 INFORMS conference. Resource management is of particular importance within healthcare because of the system's unique objectives and challenges. We provide a perspective of the current health-care literature, focusing on recent papers in planning and scheduling and reviewing them along four dimensions: (1) who or what is being scheduled, (2) the time horizon of the scheduling or planning, (3) the level of uncertainty inherent in the planning, and (4) the decision criteria. With this perspective on the literature we observe that the problems at the extreme ends of the time dimension deserve more attention: long-term planning/staffing and real-time task assignment.

13.1 Introduction

The USA spends a larger proportion of its gross domestic product on health-care expenditures than does any other country in the world. Approximately one in every six dollars of GDP is spent on healthcare in the USA [34]. In addition, the USA spends more per capita on healthcare than any other country in the world [38]. Despite these vast expenditures, the USA ranks 47th in terms of life expectancy at birth behind virtually all western European countries [36]. Life expectancy at birth in the USA (78.14 years) is nearly 4 years less than that in Japan (82.07 years) and 3 years less than Canada and Australia (81.16 and 81.53 years, respectively). In the USA, 6.3 infants die per 1000 live births putting the it behind 41 other countries including Canada (5.08 deaths), South Korea (4.29 deaths), France (3.36 deaths), and Singapore (2.3 deaths per 1000 live births), which leads the 226 listed countries [39]. Of 28 countries for which data are available, the USA is first in the percent

Jonathan Turner, Sanjay Mehrotra
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA

Mark S. Daskin
University of Michigan, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_13,

of the population that is obese (30.6%), with most western countries between 8 and 15% [40]. In Japan and South Korea, the obesity rate is roughly one-tenth that of the USA.

The availability of health-care services in the USA also lags that of many other countries. Recent statistics indicate that there are 8.1 nurses per 1000 people in the USA, compared to over 10 per 1000 people in such countries as Norway (10.3), Australia (10.7), Switzerland (10.7), the Netherlands (13.4), Ireland (14), and Finland (14.7) [37]. If availability is measured in terms of hospital beds per 1000 people, the USA (at 3.3 beds per 1000 people) lags behind much of the world, ranking 81st out of 191 countries in a recent data set [35] behind such countries as Japan (14.3), Germany (8.9), France (7.7), and Israel (6.1).

Thus, in spite of massive spending, the USA trails many countries in terms of health-care *outcomes* and in terms of *available* resources per capita. Hence, using the limited resources more effectively becomes even more important in US healthcare to improve the less-than-excellent health outcomes. Planning, scheduling, and assignment of the available resources become critical. The need for using operations research tools to address such issues has been well recognized. Fries presented a comprehensive bibliography of 188 papers that had been published in healthcare over 30 years ago [22]. He noted the then dramatic increase in papers in the field observing that "more articles were published in the first four years of this decade [the 1970s] than in the two decades preceding it." The growth has continued nearly unabated. The 2008 INFORMS (Institute for Operations Research and the Management Sciences) annual conference included 57 sessions and over 200 presentations devoted to health-care issues. Two sessions and 26 presentations focused on scheduling within healthcare. A recent issue of the *European Journal of Operational Research* includes nearly a dozen health-care-related papers, at least four of which relate in some way to planning, scheduling, and allocation issues (185(3)). Since Fries' paper, many new journals devoted to health-care management have been initiated, the most prominent of which may be *Healthcare Management Science*. Traditional operations research journals have recently devoted entire issues to healthcare (e.g., IIE Transactions 40(9), 2008). The literature in this field is truly vast and it is not possible to capture all that has been done in the available space. This chapter focuses on planning and scheduling issues in healthcare.

Even though operations research has much to contribute to planning, scheduling, and assignment problems in healthcare, its attention to date has focused excessively on a relatively narrow class of problems. While these problems are important from the operational perspective of a health-care provider, the literature generally fails to address some of the more critical problems faced by health-care institutions and by the nation. Our hope is that this chapter will stimulate additional research in these critical areas.

The remainder of the chapter is organized as follows. In Section 13.2, we outline a multi-dimensional framework for planning, scheduling, and allocation problems in healthcare. The two primary dimensions are (1) who or what is being planned for and (2) time. In addition, we discuss the impact of (3) uncertainty and (4) decision criteria on the problems being modeled. Section 13.3 provides a brief,

and necessarily incomplete, overview of the available literature on health-care planning, scheduling, and allocation problems. In Section 13.4, we present conclusions and suggestions for future work.

13.2 A Multi-dimensional Taxonomy of Health-Care Resource Management

Major dimensions of a health-care resource management problem are (1) who and what; (2) the time horizon over which the resources are being managed; (3) the level of uncertainty inherent in the planning; and (4) the decision criteria. As discussed below, these dimensions distinguish resource management problems in healthcare from those arising in manufacturing, transportation, and logistics industries.

13.2.1 Who and What of Health-Care Resource Management

At least three different *entities are simultaneously being managed* in health-care systems:

- Physical resources such as surgical theaters, emergency rooms, sterilization labs, and hospital beds
- Health-care personnel including emergency physicians; residents and interns; nurses; pharmacological support and technicians
- Patients themselves

The problem complexity will vary with constrained availability of one or more of these entities, assuming that the rest are unlimited. Consider an example of these three entities associated with scheduling a vascular surgical procedure. For the surgery to take place, a surgical suite must be available. Thus, there is a need to schedule the operating rooms and to assign them to different surgery practices (e.g., vascular surgery). Second, a group of physicians, not just the surgeons, must be available at the same time. For the surgery to occur, surgical nurses, anesthesiologists, and perhaps other specialists (for example, radiologists) must also be available at the same time, thus creating the need for more coordinated personnel scheduling. Physicians have specialty areas and it is often objectionable, if not impossible, to substitute one surgeon or physician for another. This is in sharp contrast to many manufacturing operations in which, for example, one lathe operator can readily be substituted for another. Thus, there is a need to simultaneously schedule surgeons as well as operating rooms. The necessary surgical equipments (e.g., scalpels, sutures, anesthetics, and medicines) must be in place in the surgical suite, creating scheduling demands for the sterilization lab and pharmacy. A particular patient is then assigned a time in the operating room, and space must be available for the patient in an appropriate recovery room.

Thus health-care scheduling must be done accounting for (1) the other demands placed on the time of the personnel (e.g., surgeon's clinical schedule), (2) the other demands placed on the physical resources (e.g., the operating theater), (3) demand placed by the schedule on other personnel and resources affected from the schedule, and (4) the highly uncertain nature of the processes involved as discussed later in Section 13.2.3.

13.2.2 Decision Horizon

The second dimension along which it is useful to stratify the literature is the *temporal scale* or *planning horizon* affected by the managerial decisions. These decisions take place at five strategic stages as shown in Fig. 13.1.

Fig. 13.1 Temporal dimension of decisions in healthcare

Warner [64] identifies the latter four levels (yearly to patient assignment) of temporal decision making for nurse planning and scheduling and to the best of our knowledge the operations research literature has focused on these levels only.

Long-term *planning* should also address questions of *national policy* in addition to questions about the sizing of a particular operating facility. At the national level, how many new physicians should we be training to prepare for future needs anticipating the aging of the baby boomers with the attendant increase in demand for healthcare? Who is responsible for making these decisions and are their objectives consistent with those of good public policy? Given that approximately one in six Americans lack health-care coverage today, what would be the impact on the demand for medical services in general and physicians and nurses in particular of mandated health-care coverage? How would the need for physicians break down by specialty and by region under such a plan? Are there medical needs that are largely addressed for all patients today independent of whether or not the patients currently have medical insurance while other needs may dramatically increase in demand with mandated coverage? Are there some specialties that might experience a decrease in demand if national health insurance is mandated? What will happen to the demand for expensive testing equipment if another 15–20% of the population suddenly has insurance benefits? Will we experience a 15–20% increase in demand for MRI testing or is there a significant marginal usage of MRI testing (testing that

may be prescribed now simply to ensure full utilization of the facility) that would simply be driven out of the market by the more critical (and perhaps legitimate) usage demands of the newly insured?

While these issues have been largely absent from the operations research literature, others have recognized the need for national projections of supply and demand for medical personnel. The Bureau of Labor Statistics [62] projects nursing employment and the Department of Health and Human Services [63] projects physician supply and demand by specialty. The latter report also examines usage by patient age and by the type of medical coverage the individual has. Neither report examines the impact of possible changes in health-care coverage. Also, these reports provide point estimates only of supply and demand in the case of physicians. As noted in US Department of Health and Human Services [63] "projecting demand for physician services [is difficult], where much uncertainty exists regarding the characteristics of the future healthcare system" (p. 31). Thus, there is a need for improved stochastic modeling of future supply and demand in healthcare.

Staffing refers to decisions, typically made annually, dealing with the number of personnel of each type to employ at a health-care institution. For example, a hospital must decide how many LPNs and RNs to employ, how many hospitalists to have on staff, and how many internists should be granted admitting privileges. *Scheduling* decisions, the focus of Warner's paper and of much of the operations research literature as indicated below, are made every 4–6 weeks. The key issue is to assign individual health-care workers (e.g., nurses, emergency room physicians) to shifts over the time frame in question. As discussed below, there are a myriad of hard requirements which must be satisfied and soft requirements which should be satisfied if possible. The objective function typically includes penalties for violations of the soft requirements.

The fourth stage deals with *allocation* decisions. *Allocation* refers to the need to employ temporary or traveling nurses or to use float nurses to handle unexpectedly large patient demands during particular shifts and to assign individual nurses to particular units. This allocation phase is done at the shift level, with shifts typically lasting between 8 and 12 h. Clearly, longer term staffing and scheduling decisions impact the extent to which temporary or traveling nurses need to be employed [15].

The fifth and final stage of planning deals with *assignment* of personnel to individual tasks. For example, in an intensive care unit, as patients arrive from the emergency room or from an operating room and as other patients leave for less critical medical-surgical beds over the course of a shift, how should patients be assigned and re-assigned to the nursing staff that is on duty? How many patients should be assigned to a nurse? How should new patients be assigned to rooms?

While we have discussed the temporal dimension of planning in terms of nurse personnel, several of these stages also apply to physicians, attending residents, technicians, and other medical staff, as well as the physical resources used in healthcare. For example, a hospital typically decides annually the number of hours per week to allocate to each surgical specialty. On a weekly or monthly basis, blocks of time are then assigned to the surgical practices in accordance with these overall hourly quotas. Each specialty area is then tasked with allocating time within its assigned

block to individual surgeons who then schedule particular patients for the surgery that is needed.

13.2.3 Level of Uncertainty

Demand uncertainty is significantly higher in health-care situations that in manufacturing. The stochasticity arises not only from the uncertainty regarding the *need* for any particular procedure but also from the *duration* of the procedure. To understand this issue, let us return to our vascular surgery example. Unlike many manufacturing contexts where demand forecasts with reasonable accuracy are available far in advance, demand for elective surgery is often unknown until a few weeks before surgery begins while demand for emergent surgery is often unknown until a few days or hours before surgery begins. In dealing with uncertain demand in manufacturing, we can produce goods in advance of the demand and hold them in inventory. We cannot maintain an inventory of unused time in a surgical suite. Like many other services, unused capacity is lost and cannot be inventoried in healthcare. Also, manufacturers are sometimes willing to incur lost sales if inventory costs are high, whereas to a physician, rejecting a patient's need for surgery is inconceivable.

A second source of uncertainty in healthcare arises from the actual surgical task itself. Differing patient characteristics contribute significantly to this. Wright et al. highlight the difficulties associated with predicting surgical times [67]. No two patients are identical, and sometimes it is only after the surgery has begun that a surgeon knows all that may be required. For example, one coronary artery bypass graft (CABG) surgical patient may require only one vascular graft, while a different patient may require five. One patient may experience a sudden drop in blood pressure during surgery, requiring additional interventions, while a different patient surgery may proceed normally. There are also pre- and post-surgical requirements that depend on patient characteristics. For example, one CABG patient may experience a quick recovery requiring limited post-surgical monitoring, while another patient may require dedicated post-surgical support for several hours from the surgeon.

13.2.4 Decision Criteria

Planning and scheduling decisions in healthcare are fundamentally multi-objective. While decisions in manufacturing and other service industries are driven primarily by cost minimization or profit maximization, such is not the case for the health-care problems. Whereas delays and defects in manufacturing may result in lost revenue, delays or lack of proper service in healthcare may result in loss of life. Hence, cost minimization and resource utilization decisions must consider the patient safety consequences of such decisions. While the resources and personnel being scheduled are

very expensive, and administrators and physicians naturally want to maximize the utilization of these scarce resources, slack must be incorporated in the schedules in a strategic manner to accommodate unexpected emergency demands. Other examples of multiple, often competing, objectives are patient and staff satisfaction, patient continuity of care, and educational goals in case of residents.

Warner's [64] paper is seminal for nurse scheduling. It not only outlined the temporal dimensions along which decisions are made, as discussed above, but also identified six attributes that a good (nurse) schedule should possess. These include

- Coverage or the ability of a schedule to provide the adequate number of nurses needed in each shift;
- Quality of the schedule as judged by the nursing staff working the schedule;
- Stability of the schedule or the degree to which the schedule is predictable and seems to follow prescribed guidelines;
- Flexibility of the scheduling system to handle different schedule requirements;
- Fairness of the schedule across all staff; and the
- Cost of the schedule to the hospital.

Quality, stability, and fairness issues are central to achieving better staff satisfaction. Incorporation of such requirements makes the modeling of health-care resource management problems more difficult and the models become increasingly difficult to solve.

In short, resource management in healthcare has multiple temporal stages that require coordination of multiple personnel and physical resources in a highly uncertain environment, with multiple competing objectives and requirements that are often difficult to model.

13.3 Operations Research Literature on Resource Management Decisions in Healthcare

As indicated above, the operations research literature on resource management problems in healthcare is vast. Most of it focuses on intermediate-term (4–6 weeks) scheduling problems. We do not pretend to be able to review all or even a significant portion of the literature in a short chapter. Instead, we begin by referring the reader to a number of recent review papers. Cardoen, Demeulemeester, and Belien provide a recent review of roughly 125 papers on operating room scheduling problems and models [10, 11]. Burke summarizes the state of the art in nurse scheduling [9] as do Cheang et al. [13]. Ernst et al. review staff scheduling in general, but include a section on health-care applications [19]. The reader is encouraged to consult these overview papers for a more comprehensive summary of the state of the art in scheduling.

The rest of this section is organized as follows. In Sections 13.3.1, 13.3.2, 13.3.3, and 13.3.4, we focus on scheduling problems from the perspective of who or what is being scheduled. Section 13.3.1 deals with nurse scheduling, Section 13.3.2 focuses

on scheduling other medical personnel, Section 13.3.3 summarizes patient scheduling, while Section 13.3.4 turns to facility scheduling. In Section 13.3.5, we shift our focus to examine longer term planning issues.

13.3.1 Nurse Scheduling

Nurse scheduling, or alternatively nurse rostering, is the problem of assigning nurses to shifts over a 4-to 6-week period of time. Nurse scheduling is difficult because hospitals must be staffed by nurses (and others) 24 h a day. This gives rise to satisfaction and fairness issues. Even simple scheduling problems that can be formulated as network flow problems when the planning day ends at some point in time (e.g., the store closes at 8 p.m.) become NP-hard when 24-h staffing is required. From the perspective of the nurses, poor schedules lead to numerous problems. There is evidence that increasing the patient-to-nurse ratio correlates with in-patient mortality rates and increased medication errors [20, 24, 29, 47, 52, 53, 56]. Nurses themselves may suffer adverse health impacts from poor shift schedules including peptic ulcers, coronary heart disease, and compromised pregnancy outcomes [30], colorectal cancer [50], and breast cancer [51].

Nurse schedules typically take one of three forms. A *cyclic* schedule develops a set of different cycles of work (e.g., schedule 1 might be Monday, Wednesday, and Friday 8 a.m. to 8 p.m.; schedule 2 might be Tuesday, Wednesday, Thursday, and Friday 4 p.m. to midnight plus Sunday 8 p.m. to 8 a.m.). Nurses rotate through a series of schedules in such a way that the staffing needs of all shifts are satisfied throughout the planning horizon and work rules for the nursing staff are obeyed. As an example of a violation of a typical work rule, schedule 2 could not precede schedule 1 since the nurse would then effectively have to work from Sunday at 8 p.m. through Monday at 8 p.m., a 24-h period. Note also that the two schedules above account for 36 and 40 h of work, respectively, again reflecting the fact that different nurses may be contracted for different numbers of hours of work each week and/or the nursing staff may not always work the same number of hours each week. Cyclic schedules tend to be very inflexible. Some amount of flexibility is necessary in personnel scheduling so that staff can attend to emergencies or other personal needs. With the current shortage of nursing staff across the country—a shortage that most predict will get worse before it gets better—flexible schedules are increasingly important as hospitals attempt to retain their best staff. Bard and Purnomo applied Lagrangian relaxation [6] to the cyclic scheduling problem as well as branch and price techniques [46].

At the other extreme is *self-scheduling* in which nurses sign up for individual shifts with limits on the total number permitted per shift by the hospital administration. Again, fairness can be an issue in that those who sign up early often get the preferred schedules while those who sign up late (either due to their own tendency to procrastinate or due to a monthly rotating order of signups) often get the less desirable ones. Bailyn et al. [7] describe a recent effort in implementing such approaches at a unit consisting of 70 RNs.

Preference scheduling allows the medical staff (e.g., the nurses) to express preferences for specific shifts during the planning horizon. Preferences may be either positive, indicating that the individual wants to work at that time, or negative, indicating that they do not want to be on duty then. The objective is then to maximize the total staff preferences, perhaps in combination with penalties for violating some of the soft constraints. Recent efforts to automate generation of nurse schedules based on mathematical modeling approaches while incorporating nurse preference are described in Rönnberg and Larsson [48]. The responses from the nurses in their pilot study were both expected and skeptical; expected because of the time-consuming work and difficulties associated with the manual process and skeptical mainly because of the nurses' loss of influence on the outcome of the scheduling. These authors conclude, "Because of the nurses' scepticism it is important to emphasize that the optimization tool only provides a qualified suggestion for a schedule, and encourage the nurses to make minor adjustments themselves if beneficial." An alternative to mathematical programming approach as a means of dealing with nurse preferences is the use of auctions [17].

Bard and Purnomo identify 13 different categories of constraints that typically appear in nurse scheduling problems and classify them as either hard constraints—those that must be satisfied in any schedule—or soft constraints—those that should ideally be satisfied but whose violation is penalized in the objective function [3]. Most other authors (e.g., Burke et al. [8, 9], Wright et al. [68], Parr and Thompson [43]) also make such distinctions. A hard constraint might stipulate that there must be an 8-h break between every shift. Another might stipulate that each nurse must work at least a given number of hours each week, with the number of hours varying by nurse according to the type of contract they have with the hospital. Another hard constraint might be that a nurse cannot work more than six consecutive days. A related soft constraint might penalize the objective function if a nurse works six consecutive days in a row as the goal might be to work no more than five consecutive days.

Bard and Purnomo address the shorter term allocation problem or reactive scheduling problem as they term it [4]. The problem is formulated as a mixed integer programming model and is solved with up to 200 nurses. Bard and Purnomo examine the problem using column generation [3–5]. The models in Burke et al. [8, 9] were solved using an evolutionary and a neighborhood search heuristic. Wright et al. [68] develop a bi-criteria scheduling model and perform computational experiments to evaluate how mandatory nurse-to-patient ratios and other policies impact a schedule's cost and desirability (from the nurses' perspective). Their findings suggest that (i) the nurse wage costs can be highly nonlinear with respect to changes in mandatory nurse-to-patient ratios of the type being considered by legislators; (ii) the number of undesirable shifts can be substantially reduced without incurring additional wage costs; (iii) more desirable scheduling policies, such as assigning fewer weekends to each nurse, have only a small impact on wage costs; and (iv) complex policy statements involving both single-period and multi-period service levels can sometimes be relaxed while still obtaining good schedules that satisfy the nurse-to-patient ratio requirements. Thompson and Parr [59] consider a multi-objective

nurse scheduling problem using a weighted sum cost function and use a simulated annealing-based heuristic.

Mullinax and Lawley [32] develop an integer linear programming model that assigns patients to nurses in a neonatal intensive care unit. The nurseries are divided into a number of physical zones. The authors used a zone-based heuristic that assigns nurses to zones and computes patient assignments within each zone. Earlier approaches and patient classification systems ignore uncertainty. Punnakitikashem et al. [45] present a stochastic integer programming model for the nurse assignment problem. This model is further integrated into the staffing problem in Punnakitikashem [44]. The objective is to minimize excess workload for nurses. By considering randomness in the models, Punnakitikashem [44] shows that better staffing and assignments decisions are possible.

13.3.2 Scheduling of Other Health-Care Professionals

The literature on scheduling problems for other health-care professionals is limited. Scheduling of emergency room physicians was considered in Carter and Lapierre [12] and subsequently discussed in Gendreau et al. [23]. The types of constraint used in these papers are similar to those considered in the nurse scheduling/rostering problem. The latter paper discusses the use of column generation, tabu search, and constraint logic programming as possible methods for solving this problem. Constraint logic programming was also used recently in Edqvist [18] for physician scheduling in a clinical setting.

The resident scheduling problem is considered in Ozkarahan [42], Sherali et al. [54], White and White [66], Day et al. [16], and Topaloglu [60, 61]. The educational benefit of the activities in which residents and interns are engaged [1, 2] is a critical factor in resident and intern scheduling that is absent from nurse scheduling. Only one of these papers [16] deals with the educational facets of these problems.

13.3.3 Patient Scheduling

Patient scheduling, in contrast to nurse and resident scheduling, typically operates at the daily time frame. The key issues in patient scheduling revolve around the need to minimize (1) the physician's idle time, (2) the time at which the physician sees the last patient of the day, and (3) the total patient waiting time. Thus, this problem is also multi-objective. These times are impacted not only by the scheduled appointment times—the key decision variables in most of the literature—but also by the variability in arrival times about the scheduled time, the mean and variability of the service time(s), and the patient no-show rates. Many of the commonly used scheduling rules, including the well-known Bailey–Welch rule [65], seem to give priority to the physician-related times as opposed to the patient waiting time. This rule suggests scheduling two patients at the beginning of the day and scheduling successive patients at intervals of one mean service time following the initial appointment time.

Clearly, by ignoring variability in the service times and by initializing the system with one waiting patient, this model prioritizes physician idle times more highly than patient delays.

Murthuraman and Lawley consider the problem of dynamically assigning patients to appointment slots at an outpatient clinic during a day accounting for no show rates [33]. Liu, et al. also studied the impact of patient no-shows and cancellations on outpatient scheduling [31]. They determine optimal scheduling policies using Markov decision processes comparing them to an open access (OA) policy. OA policies do not schedule patients very far in advance and are favored by patients who walk in. Ho and Lau analyzed a variety of factors impacting outpatient scheduling and found that the Bailey–Welch rule is remarkably robust with respect to a range of input conditions and assumptions [27]. Kaandorp and Koole [28] also validate the Bailey–Welch rule by formulating a discrete time queueing model to minimize a weighted sum of patient waiting time, physician idle time, and tardiness.

13.3.4 Facility Scheduling

Facility scheduling is intimately linked to patient scheduling in many cases, as it is the patients who are being scheduled for services at a facility. One additional facet of facility scheduling that is typically not present in scheduling physicians at a clinic is the need to balance scheduled patient needs with those of unscheduled emergency patients. Green et al. analyze scheduling problems for a facility that handles inpatients, outpatients, and emergency cases, each with different arrival characteristics, service needs, and delay costs [25]. Swisher et al. adopt a long-term perspective and use simulation coupled with experimental design to analyze the impact of facility design on patient delays and facility costs [57]. Factors considered in the design of the facility include the number of physician assistants, nurses, and medical assistants, as well as the number of check-in rooms, examination rooms, and specialty rooms.

Of all the facilities in a health-care institution, operating rooms have attracted the most attention within the scheduling literature. Cardoen et al. [11] provide an updated review of operating room scheduling literature focusing on manuscripts published in or after 2000, which make up around half of the manuscripts they found. They label each of these papers according to nine different fields, some of which are methodology, whether the models include stochasticity, and the level of the decision (date, time, room, or capacity). The vast majority of the papers reviewed by Cardoen et al. focus on patient level decision making, discipline level (as in vascular surgery, cardiology, etc.), or other levels. Most also focus on date, time, or room decisions rather than budgeting capacity. One of these is by Denton and Gupta [26] and slots patients for operating room times using a stochastic program. Not all operating room scheduling literature is focused on the short- to medium-term problem, however. O'Neill and Dexter [41] show how a hospital can analyze population data and federal surgical rates to plan for operating room capacity needs by surgical specialty. Others, like Testi et al. [58], bridge decision making across

short- and long-term planning horizons. They use optimization and simulation models to budget operating room capacity on a weekly basis, allocate operating room time to surgical units, and assign patients to time slots.

13.3.5 Longer Term Planning

Longer term personnel planning issues have also been considered. Sinreich and Jabali [55] focus on the task of determining patient demand and finding work shift schemes for meeting this demand. They propose a staggered work shift schedule, each starting on the hour, to better match the demand on different emergency room resources. Using a simulation–optimization framework and an iterated approach to schedule physicians and nurses they show a significant reduction in the required physician (8–18%), nurse (13–47%), and technician (3–33%) hourly capacity, while maintaining the current patient length of stay operational measures.

Franz and Miller [14, 21] formulate an integer programming problem for assigning residents to rotations over the course of a year recognizing many of the critical constraints that limit the possible assignments. Interestingly, although the [21] project was successful in terms of the ability of the model to solve the problem at hand, overall it was deemed unsuccessful. The authors state (p. 277), "the implementation effort must be regarded as a failure at this time." They cite six primary causes for the failure, including

1. the senior management who could direct the implementation of the model were not directly involved in the scheduling task and as such had no vested interest in altering the status quo;
2. the residents tasked with scheduling are physicians and scheduling is not their primary concern;
3. the group doing scheduling in a year changes to a different group in the following year resulting in a lack of continuity;
4. the person who championed the original effort had left the hospital;
5. the confidence of the residents in the automated process inhibits their willingness to forgo the more transparent manual assignment process, and
6. the residents believe, despite the available evidence, that they can do better than a computer.

We emphasize that the problems listed above are likely to plague any scheduling or planning effort if they are not continuously addressed throughout the modeling, analysis, and implementation processes. It is therefore important to share and benefit from lessons such as those reported by Cohn et al. [14], who report a more positive experience with their effort to schedule psychiatric medical residents for the Boston University School of Medicine. They attribute their success to (1) on-going communication between their team and the application expert which corrected their earlier mistakes in problem formulations, (2) not striving for "optimality" as the objective

but presenting "acceptable" solution choices to the application expert, (3) modifying smaller models rather than focusing on solving one single larger model, (4) the speed for providing solutions, instead of getting bogged down into technical research questions.

Long-term facility planning has received some, though relatively little, attention in the operations research literature. Santibáñez et al. [46] report on the use of an integer programming model to plan for future clinical practices across a 12 hospital system in Canada. The model assumes that patients can be assigned to hospitals, whereas many patient/hospital assignments are the result of patient choices and not system-wide allocation rules. While no final decisions have been made based on the model, the authors report that they were "successful in that the configurations we analyzed in this planning initiative were useful and relevant to executive management in developing a hospital configuration plan" (p. 206).

13.4 Summary, Conclusions, and Directions for Future Work

In Section 13.2, we outlined a multi-dimensional framework for examining planning, staffing, scheduling, allocation, and assignment problems in healthcare. The first dimension dealt with who or what was being scheduled: (a) healthcare providers including nurses, residents, and physicians; (b) medical facilities; and (c) patients. The second dimension dealt with the planning horizon over which the scheduling decisions were relevant: (a) long-term planning, (b) annual staffing, (c) intermediate-term or monthly scheduling problems, (d) short-term allocation problems, and (e) real-time task allocation issues. Uncertainty and the decision criteria were identified as two additional facets of health-care planning and scheduling that complicate the problems at hand. In Section 13.3, we provided a sampling of the available literature, recognizing that a complete analysis is beyond the scope of this (or any) chapter.

The vast majority of the operations research literature seems to focus on intermediate-term provider scheduling problems and short-term patient and facility scheduling. Provider scheduling problems focus on ensuring adequate coverage of each shift during a month accounting for work rule restrictions (hard constraints) and employee preferences (often modeled as soft constraints). Short-term and facility scheduling models try to balance the costs of physician idle time and daily task completion times with the costs of patient waiting time.

Two directions for research seem to emerge from this review. First, many of the most critical health-care problems facing the country today are not related to short-term or intermediate-term scheduling. Instead, they deal with long-term planning decisions. As indicated above, if mandatory health insurance is adopted nationally, there are likely to be significant *but differential* impacts on provider institutions and the demands placed on health-care facilities. Also, as the population continues to age, additional demands will be placed on limited facilities and already overworked personnel. The resolution of these problems is not likely to lie in (marginally)

improved schedules for outpatients or diagnostic facilities. Thus, some research attention should be devoted directly to these issues.

At the same time, relatively little research seems to have focused on real-time task assignment problems. These problems are ripe for stochastic optimization in which a decision must be made, for example, about which nurse to assign to a new patient entering an intensive care unit, accounting for unknown departure times of current patients and the arrivals of additional patients. Similarly, real-time rescheduling of operating rooms in response to unexpected delays, or early terminations of procedures, is also an area for potential research.

Finally, there seems to have been relatively little work that cuts across the two primary dimensions of planning and scheduling. In particular, long-term planning should be influenced by the best practices in intermediate-term (monthly) scheduling. Inefficient monthly schedules are likely to result in long-term facility sizing and employee hiring decisions that incur excess cost. The best models for long-term annual planning, however, are not likely to include an embedded shift-scheduling model as this will represent excessive detail. Instead, some good method of approximating the monthly scheduling costs should be developed for long-term scheduling problems.

References

1. ACGME (2007) Duty hours language. http://www.acgme.org/acWebsite/dutyhours/dh_index.asp
2. ACGME (2008) The ACGME's approach to limit resident duty hours 2007-2008: A summary of achievements for the fifth year under the common requirements. www.acgme.org/acWebsite/dutyhours/dh_achievesum0708.pdf
3. Bard JF, Purnomo HW (2005) Hospital-wide reactive scheduling of nurses with preference considerations. IIE Transactions 37(7):589–608
4. Bard JF, Purnomo HW (2005) Preference scheduling for nurses using column generation. European Journal of Operational Research 164(2):510–534
5. Bard JF, Purnomo HW (2005) A column generation-based approach to solve the preference scheduling problem for nurses with downgrading. Socio-Economic Planning Sciences 39(3):193–213
6. Bard JF, Purnomo HW (2007) Cyclic preference scheduling of nurses using a Lagrangian-based heuristic. Journal of Scheduling 10(1):5–23
7. Bailyn L, Collins R, et al (2007) Self-scheduling for hospital nurses: An attempt and its difficulties. Journal of Nursing Management 15:72–77
8. Burke EK, Causmaecker PD, et al (2002) A multi criteria meta-heuristic approach to nurse rostering. Proceedings of the Evolutionary Computation Congress, IEEE Press, Honolulu, HI
9. Burke EK, Causmaecker PD, et al (2004) The state of the art of nurse rostering. Journal of Scheduling 7:441–499
10. Cardoen B, Demeulemeester E, et al (2007) Operating room planning and scheduling: A literature review. Department of Decision Sciences and Information Management, Faculty of Business and Economics, Catholic University, Leuven
11. Cardoen B, Demeulemeester E, et al (2009) Operating room planning and scheduling: A literature review. European Journal of Operational Research, 201(3):921–932

12. Carter MW, Lapierre SD (2001) Scheduling emergency room physicians. Healthcare Management Science 4:347–360
13. Cheang B, Li H, et al (2003) Nurse rostering problems—A bibliographic survey. European Journal of Operational Research 151:447–460
14. Cohn A, Root S, et al (2006) Using mathematical programming to schedule medical residents. Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109
15. Davis A (2009) The nurse shortage crisis. IE 490, Operations Research in Healthcare, Term Project, Evanston, IL. Department of Industrial Engineering, Northwestern University
16. Day TE, Napoli JT, et al (2006) Scheduling the resident 80-hour work week: An operations research algorithm. Current Surgery 63(2):136–142
17. De Grano ML, Medeiros DJ, et al (2008) Accommodating individual preferences in nurse scheduling via auctions and optimization. Healthcare Management Science forthcoming
18. Edqvist S (2008) Scheduling physicians using constraint programming. Master Thesis in Engineering Physics, Report UPTEC F08 064, Faculty of Science and Technology, Uppsala University, Sweden
19. Ernst AT, Jiang H, Krishnamoorty M, Sier D (2004) Staff scheduling and rostering: A review of applications, methods and models. European Journal of Operational Research 153:3–27
20. Fitzpatrick JM, While AE, et al (1999) Shift work and its impact upon nurse performance: Current knowledge and research issues. Journal of Advanced Nursing 29(1):18–27
21. Franz LS, Miller JL (1993) Scheduling medical residents to rotations—Solving the large-scale multiperiod staff assignment problem. Operations Research 41(2):269–279
22. Fries BE (1976) Bibliography of operations research in healthcare systems. Operations Research 24(5):801–814
23. Gendreau M, Ferland J, et al (2006) Physician scheduling in emergency rooms. Proceedings of the 6th international conference on Practice and theory of automated timetabling VI, Brno, Czech Republic, pp. 53–66
24. Gold DR, Rogacz S, et al (1992) Rotating shift work, sleep, and accidents related to sleepiness in hospital nurses. American Journal of Public Health 32(7):1011–1015
25. Green LV, Savin S, et al (2006) Managing patient service in a diagnostic medical facility. Operations Research 54(1):11–25
26. Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. IIE Transactions http://www.informaworld.com/smpp/title∼db=all∼content=t713772245∼tab=issueslist∼branches=40 - v40 40(9):800–819
27. Ho CJ, Lau HS (1992) Minimizing total-cost in scheduling outpatient appointments. Management Science 38(12):1750–1764
28. Kaandorp GC, Koole G (2007) Optimal outpatient appointment scheduling. Healthcare Management Science 10:217–229
29. Kane RL, Shamliyan T, et al (2007) Nurse staffing and quality of patient care. Evidence Report/Technology Assessment, Number 151, Minnesota Evidence-based Practice Center, Minneapolis, MN
30. Knutsson A (2003) Health disorders of shift workers. Occupational Medicine 53: 103–108
31. Liu N, Ziya S, et al (2008) Dynamic scheduling of outpatient appointments under patient no-shows and cancellation. Chapel Hill, NC, Department of Statistics and Operations Research, University of North Carolina
32. Mullinax C, Lawley M (2002) Assigning patients to nurses in neonatal intensive care. Journal of Operational Research Society 53(1):25–35
33. Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Transactions 40(9):820–837
34. Nationmaster.com (2009a) Health statistics—Expenditures per capita—Current US$ (most recent) by country
35. Nationmaster.com (2009b) Health statistics—Hospital beds—per 1,000 people (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_hos_bed_per_1000_peo-beds-per-1-000-people

36. Nationmaster.com (2009c) Health statistics—Life expectancy at birth—Total population (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_lif_exp_at_bir_tot_pop-life-expectancy-birth-total-population
37. Nationmaster.com (2009d). Health statistics—Nurses (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_nur-health-nurses
38. Nationmaster.com (2009e) Health statistics—Expenditures, total—% of GDP (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_exp_tot_of_gdp-health-expenditure-total-of-gdp
39. Nationmaster.com (2009f) Heath statistics—Infant mortality rate—Total (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_inf_mor_rat_tot-health-infant-mortality-rate-total
40. Nationmaster.com (2009g) Heath statistics—Obesity (most recent) by country. Retrieved June 11, 2009, from http://www.nationmaster.com/graph/hea_obe-health-obesity
41. O'Neill L, Dexter F (2007) Tactical increases in operating room block time based on financial data and market growth estimates from data envelopment analysis. Anesthesia and Analgesia 104(2):355–368
42. Ozkarahan I (1994) A scheduling model for hospital residents. Journal of Medical Systems 18(5):251–265
43. Parr D, Thompson JM (2007) Solving the multi-objective nurse scheduling problem with a weighted cost function. Annals of Operations Research 155(1):279–288
44. Punnakitikashem P (2007) Integrated nurse staffing and assignment under uncertainty, Ph.D. Thesis, University of Texas at Arlington, Arlington, TX
45. Punnakitikashem P, Rosenberger JM, et al (2008) Stochastic programming for nurse assignment. Computational Optimization and Applications 40(3):321–349
46. Purnomo HW, Bard JF (2007) Cyclic preference scheduling for nurses using branch and price. Naval Research Logistics 54(2):200–220
47. Rogers AE, Hwang WT, et al (2004) The working hours of hospital staff nurses and patient safety. Health Affairs 23(4):202–212
48. Rönnberg E, Larsson T (2010) Automating the self-scheduling process of nurses in Swedish healthcare: a pilot study. Health Care Management Science 13(1):35–53
49. Santibáñez P, Bekiou G, Yip K (2009) Fraser Health uses mathematical programming to plan its inpatient hospital network. Interfaces 39(3):196–208
50. Schernhammer ES, Laden F, et al (2003) Night-shift work and risk of colorectal cancer in the nurses' health study. Journal of National Cancer Institute 95(11):825–828
51. Schernhammer ES, Laden F, et al (2001) Rotating night shifts and risk of breast cancer in women participating in the nurses' health study. Journal of National Cancer Institute 93(20):1563–1568
52. Seago JA (2008) Nurse staffing, models of care delivery, and interventions. http://www.ahrq.gov/clinic/ptsafety/chap39.htm
53. Seago JA, Williamson A, et al (2006) Longitudinal analyses of nurse staffing and patient outcomes: More about failure to rescue. JONA The Journal of Nursing Administration 36(1):13–21
54. Sherali H, Ramahi M, et al (2002) Hospital resident scheduling problem. Production Planning and Control 13(2):220–233
55. Sinreich D, Jabali O (2007) Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. Health Care Management Science 10:293–308
56. Stanton MW, Rutherford M (2004) Hospital nurse staffing and quality of care. Agency for Healthcare Research and Quality, Rockville, MD
57. Swisher JR, Jacobson SH, et al (2001) Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. Computers & Operations Research 28(2):105–125
58. Testi A, Tanfani E, et al (2007) A three-phase approach for operating theatre schedules. Healthcare Management Science 10:163–172

59. Thompson JM, Parr D (2007) Solving the multi-objective nurse scheduling problem with a weighted cost function. Annals Operations Research 155(1):279–288
60. Topaloglu S (2006) A multi-objective programming model for scheduling emergency medicine residents. Computers and Industrial Engineering 51(3):375–388
61. Topaloglu S (2008) A shift scheduling model for employees with different seniority levels and an application in healthcare. European Journal of Operation Research, 198(3):943–957
62. US Bureau of Labor Statistics (2009) Occupational outlook handbook, 2008-2009 edition, http://www.bls.gov/oco/pdf/ocos083.pdf
63. US Department of Health and Human Services (2006) Physician supply and demand: Projections to 2020. HRSA, Bureau of Health Professionals http://bhpr.hrsa.gov/healthworkforce/reports/physiciansupplydemand/default.htm
64. Warner DM (1976) Scheduling nursing personnel according to nursing preference: A mathematical programming approach. Operations Research 24(5):842–856
65. Welch J, Bailey N (1952) Appointment systems in hospital outpatient departments. The Lancet 1:1105–1108
66. White CA, White GM (2003) Scheduling doctors for clinical training unit rounds using tabu optimization. LNCS: Practice and theory of automated timetabling IV, Springer, Berlin, pp. 120–128
67. Wright IH, Kooperberg C, et al (1996) Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates. Anesthesiology 85(6):1235–1245
68. Wright PD, Bretthauer KM, et al (2006) Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages. Decision Sciences 37(1):39–70

Chapter 14
Optimizing Happiness

Manel Baucells, Rakesh K. Sarin

Abstract We consider a resource allocation problem in which time is the principal resource. Utility is derived from time-consuming leisure activities, as well as from consumption. To acquire consumption, time needs to be allocated to income generating activities (i.e., work). Leisure (e.g., social relationships, family, and rest) is considered a basic good, and its utility is evaluated using the Discounted Utility Model. Consumption is adaptive and its utility is evaluated using a reference-dependent model. Key empirical findings in the happiness literature can be explained by our time allocation model. Further, we examine the impact of projection bias on time allocation between work and leisure. Projection bias causes individuals to overrate the utility derived from income; consequently, individuals may allocate more than the optimal time to work. This misallocation may produce a scenario in which a higher wage rate results in a lower total utility.

14.1 Introduction

> "The constitution only gives you the right to pursue happiness. You have to catch it yourself."
>
> — Benjamin Franklin

The Ancient Greeks believed that happiness was controlled by luck, fate, or the gods and was beyond human control [38]. Socrates and Aristotle regarded the human desire to be happy as self-evident and focused instead on how to become happy. In recent years, the science of happiness has emerged as a new area of research that attempts to determine what makes us happy. This area of research has at its foundation the measurement of happiness or well-being by means of self-reports.

Manel Baucells
Department of Managerial Decision Sciences, IESE Business School, Barcelona, Spain

Rakesh K. Sarin
Decisions, Operations & Technology Management Area, UCLA Anderson School of Management, University of California, Los Angeles. CA, USA

M.S. Sodhi, C.S. Tang (eds.), *A Long View of Research and Practice in Operations Research and Management Science*, International Series in Operations Research & Management Science 148,
DOI 10.1007/978-1-4419-6810-4_14,

In line with Easterlin [16] and Frey and Stutzer [21], we use the terms happiness, well-being and life satisfaction interchangeably and assume that these measures are a satisfactory empirical approximation of individual utility.

In developed countries, particularly in the United States, economic progress is a key factor in improving individuals' well-being. Tocqueville [55] observed, "The lure of wealth is therefore to be traced, as either a principle or an accessory motive at the bottom of all that the Americans do, this gives to all their passions a sort of family likeness." Survey results show, however, that happiness scores have remained flat in developed countries despite considerable increases in average income. In Japan, for example, a fivefold increase in real per capita income has led to virtually no increase in average life satisfaction (Figure 14.1). A similar pattern holds for the United States and Britain. In spite of these survey results, we contend that most people believe that more money will buy them more happiness.

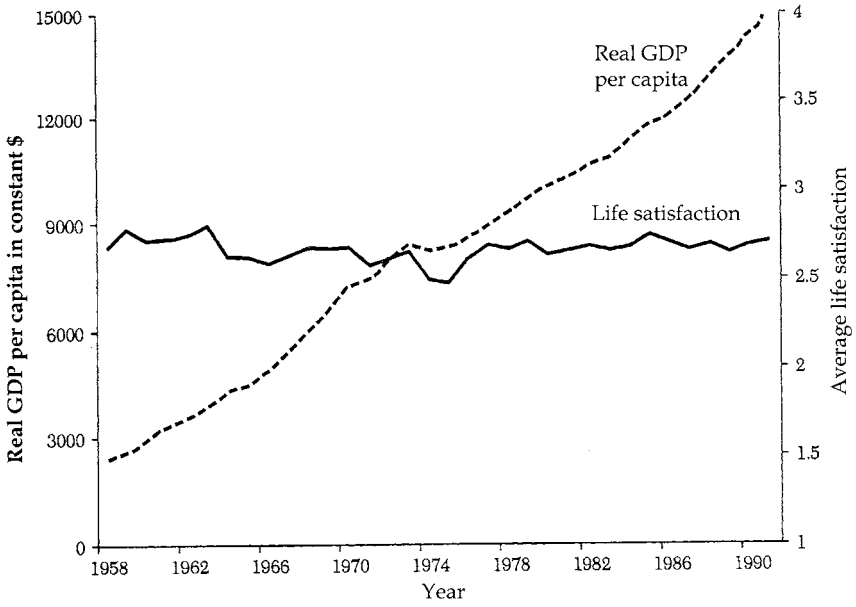


Fig. 14.1 Satisfaction with life and income per capita in Japan between 1958 and 1991. *Source:* [21, figure 2]

The purpose of this chapter is twofold. The first is to show that an adaptation and social comparison model of time allocation is consistent with key empirical findings on the relationship between money and happiness. The second is to show that under the plausible psychological assumption of projection bias there could be a misallocation of time resulting in some paradoxical predictions. It is because of projection bias that individuals believe that more money will buy them a lot more happiness than it actually does, and this may even lead to a scenario in which a higher wage rate results in a lower total utility.

We present our adaptation and social comparison model of time allocation in Section 14.2. An individual allocates a fixed amount of time between work and leisure in each period. The total utility is the discounted sum of utility derived from consumption and leisure. Leisure (e.g., time spent with friends and family) provides direct utility and is not adaptive. In contrast, there is evidence in the literature that beyond a set level of income at which basic needs are met, consumption is adaptive. The carrier of per-period utility of consumption is therefore the relative consumption with respect to a reference level. In general, the reference level of consumption depends on past consumption and social comparison. A rational individual will allocate the same fixed proportion of time to work and leisure in each period (say 40% to work and 60% to leisure) and choose an increasing consumption path over time.

In Section 14.3, we summarize some key empirical findings from the "happiness" literature. Our model, under the assumption of optimizing individual utility, is consistent with some of the findings in the literature. Our model can explain (1) why happiness scores in developed countries are flat in spite of considerable increases in average income and (2) why there is a positive relationship between individual income and happiness within a society at any given point in time. However, this optimization model cannot explain, without some further assumptions, the puzzle: *Why do we believe that more money will buy us lot more happiness than it actually does?*

In Section 14.4, we introduce projection bias into our model. Projection bias causes people to underestimate the effects of adaptation, which in turn causes them to overestimate the utility derived from adaptive goods. This is akin to buying more food at the grocery store when hungry or ruling out the possibility of a large turkey dinner for Christmas after finishing a hearty meal at Thanksgiving. Similarly, an individual who moves to a more prosperous neighborhood may insufficiently account for the increased desire for fancy cars and a higher standard of living that will occur once he begins to compare himself to and identify with his new neighbors. A pernicious effect of projection bias may be that an individual continues to allocate more and more time to work at the expense of leisure.

In Section 14.5, we examine the impact of wage rate on total utility. Under projection bias, an individual may allocate a greater amount of time to work than what is optimal. The resulting misallocation of time between work and leisure could actually lower total utility at higher wage rates.

Social comparison has been found to be a determinant of behavior in both human and animal studies. In Section 14.6, we examine the implications of our model when reference levels are influenced by social comparison.

An underlying tenet of our human condition is that to gain happiness, you must either earn more or desire less. Indeed, in our model, initial adaptation level and social comparison act to reduce the available budget. Reference levels can be moderated through reframing or perspective seeking activities. Such activities, however, require an investment of time. In Section 14.7, we extend the time allocation model to include the possibility that reference levels can be influenced by investing time in reframing activities such as meditation or other spiritual practices.

Finally, we conclude our chapter in Section 14.8 and discuss some implications of our model to improve individual and societal well-being.

14.2 Time Allocation Model

We consider a simple model of work–leisure decisions. In each period t, $t = 1$ to T, an individual divides one unit of time between work, w_t, and leisure, ℓ_t. Work produces income at a rate of μ units of money per unit of time spent at work. For simplicity, this wage rate is constant over the T periods. The individual anticipates the total amount of income generated by work during the entire planning horizon $(\mu \sum_{t=1}^{T} w_t)$ and plans consumption, c_t, $t = 1$ to T, so that total consumption $(\sum_{t=1}^{T} c_t)$ does not exceed total income. For simplicity, we assume that the individual borrows and saves at an interest rate of zero percent. We also set the price of the consumption good to a constant over time that is equal to one unit.

The individual derives utility from both consumption (i.e., necessities and conveniences of life) and leisure (e.g., time spent with friends and family, active and passive sports, rest). We assume that the per-period utility derived from consumption and leisure is separable and that the total utility is simply the discounted sum of per-period utilities.

We posit that leisure provides direct utility and is not reference dependent. One always enjoys time spent with friends and family. Sapolsky et al. [47] observed that amongst the baboons of the Serengeti, those who had more friends suffered from less stress (measured by levels of stress hormones including cortisol). Cicero said, "If you take friendship out of life, you take the sun out of the world." Similarly, family warmth, sleep, sex, and exercise improve life satisfaction. Some aspects of leisure could indeed be adaptive, but Frank [19] argues that conspicuous consumption is much more adaptive than leisure. Leisure is often consumed more privately and is valued for itself and not often sought for the purpose of achieving prestige or status. Solnick and Hemenway [53] found that vacation days are not reference dependent. Similarly, consumption of basic goods (food and shelter) is not adaptive. Since a large part of consumption in affluent societies is adaptive, we assume for simplicity that consumption is reference dependent, but that leisure is not. Our results should hold with the weaker assumption that consumption is more reference dependent than leisure.

There is considerable evidence that the utility derived from consumption depends primarily on two factors: (1) adaptation or habituation to previous consumption levels and (2) social comparison to a reference or peer group [6, 8, 16–20, 32].

A woman who drives a rusty old compact car as a student may find temporary joy upon acquiring a new sedan when she lands her first job, but she soon adapts to driving the new car and assimilates it as a part of her lifestyle. Brickman et al. [6] find that lottery winners report only slightly higher levels of life satisfaction than the control group just a year after their win (4.0 versus 3.8 on a 5-point scale). Clark [8]

finds evidence that job satisfaction—a component of well-being—is strongly related to changes in pay, but not levels of pay. Klein [30] reports that when monkeys were offered raisins and not the customary apple, their neurons fired strongly in response to the welcome change. After a few repetitions, this euphoria stopped as the animals had adapted to the better food. People also adapt to country clubs and dining in fine restaurants. A crucial implication of adaptation is that the utility derived from the same $3,000 per month worth of consumption is quite different for someone who is used to consuming that amount of goods and services than for someone who is used to consuming only $2,000 per month. Several authors have proposed models that account for adaptation in the determination of the total utility of a consumption stream [42, 45, 59, 60].

In addition to adaptation, the utility derived from consumption also depends on the consumption of others in an individual's peer group. Driving a new Toyota sedan when everyone else in the peer group drives a new Lexus sedan seems quite different than if others in the peer group drive economy cars. Frank [18, 19] provides evidence from the psychological and behavioral economics literature that well-being or satisfaction depends heavily on social comparison. Solnick and Hemenway [53, table 2] asked students in the School of Public Health at Harvard to choose between living in one of two imaginary worlds in which prices are the same. In the first world, you get $50,000 a year, while other people get $25,000 a year (on average). In the second world, you get $100,000 a year, while other people get $250,000 a year (on average). A majority of students chose the first world.

People are likely to compare themselves to those who are similar in income and status. A university professor is unlikely to compare herself to a movie star or a homeless person. She will most likely compare her lifestyle to those of other professors at her university and similarly situated colleagues at other, comparable universities. Medvec et al. [39] find that Olympic bronze medalists are happier than Olympic silver medalists, as the former compare themselves to the athletes who got no medal at all, whereas the latter have regrets of missing the gold.

Relative social position influences biochemical markers such as serotonin in vervet monkeys [37]. When a dominant monkey is placed in an isolation cage, a new monkey rises to the dominant position. The serotonin level increases in the newly dominant monkey and decreases in the formerly dominant monkey. Elevated levels of serotonin are found in the leaders of college fraternities and athletic teams. Higher concentrations of serotonin are associated with better mood and enhanced feelings of well-being.

We now state our adaptation and social comparison model of time allocation. We assume the discount factor to be 1. The set of decision variables in our model comprises three vectors, each with T components. The first vector is leisure, $\mathbf{l} = (\ell_1, \ell_2, \ldots, \ell_T)$, measured in time units. The second vector is work, $\mathbf{w} = (w_1, w_2, \ldots, w_T)$, also measured in time units. The third vector is consumption, $\mathbf{c} = (c_1, c_2, \ldots, c_T)$, measured in dollars. All three vectors take non-negative values. The individual's total utility, interpreted as happiness or life satisfaction, is given by

$$V(\mathbf{l},\mathbf{c}) = \sum_{t=1}^{T} u(\ell_t) + \sum_{t=1}^{T} v(c_t - r_t), \tag{14.1}$$

$$r_t = \sigma s_t + (1-\sigma) a_t,\ t = 1,\ldots,T, \tag{14.2}$$

$$a_t = \alpha c_{t-1} + (1-\alpha) a_{t-1},\ t = 2,\ldots,T, \tag{14.3}$$

where a_1 and s_t, $t = 1,\ldots,T$, are given.

In the above model, r_t is the reference level in period t. The reference level is a convex combination of social comparison level, s_t, and adaptation level, a_t. The adaptation level is the exponentially weighted sum of past consumptions in which recent consumption levels are given greater weight than more distant past consumption levels.

For the remainder of the chapter, the initial adaptation level, a_1, will be set to zero by default. Both u and v are normalized to take a value of zero if evaluated at zero. The first component, u, is the contribution of leisure to happiness; the second component, v, is the contribution of consumption to happiness. Both u and v are concave and twice differentiable. To capture the phenomenon of loss aversion [28, 56], we allow v to be non-differentiable at zero, with $v'(0^-) \geq v'(0^+)$.[1] Loss aversion is an important feature of adaptation models, as it imparts the behavioral property that the individual will be reluctant to choose negative values for the argument of v—that is, to choose consumption below the adaptation level (see Figure 14.2).

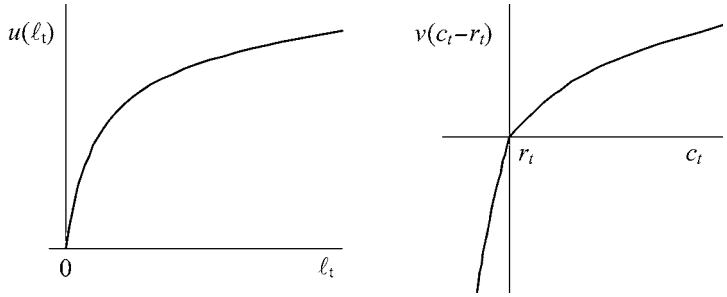


Fig. 14.2 Exemplary per-period utility for leisure and consumption

That leisure is considered a basic good implies that the per-period utility of leisure depends solely on the leisure time experienced during that period. For basic goods, the Discounted Utility Model is appropriate [4]. In contrast to leisure, consumption is considered an adaptive good. It contributes positively to happiness during a given period only if consumption is above some reference point; consumption

[1] It is appropriate to think of v as the value function of prospect theory. This function is usually taken to be concave for gains and convex for losses. As our focus is on the positive region of v, we assume for mathematical tractability that v is concave throughout. Empirical evidence shows that v is close to linear in the negative domain [1], so that the assumption of concavity for gains and linearity for losses is not farfetched.

below the reference point yields unhappiness. The dynamics of the adaptation level, a_t, are endogenously determined by the individual's own behavior. Specifically, the adaptation level is a convex combination of past consumption and past adaptation level [3, 59]. The parameter α measures the speed of adaptation. If $\alpha = 0$, then the reference level does not change and consumption is a basic good (for example, food and shelter in poor countries). If $\alpha = 1$, then the reference level is always equal to the previous period's consumption (e.g., buying a car in the next period that is worse than the current car would feel like a loss). For mathematical tractability and insight, we will often set $\alpha = 1$ in our examples.

Work does not contribute to utility, but does provide the budget to purchase consumption. An individual can plan consumption based on their total lifetime income. As there is just one unit of time available per period, time spent at work reduces the available time for leisure. Work yields μ monetary units per unit of time. With this in mind, the individual faces the following obvious time and money constraints:

$$\ell_t + w_t \leq 1, \quad t = 1, \ldots, T, \quad \text{and} \tag{14.4}$$

$$\sum_{t=1}^{T} c_t \leq \mu \sum_{t=1}^{T} w_t. \tag{14.5}$$

14.2.1 Optimal Allocation

The goal is to choose $(\mathbf{l}, \mathbf{w}, \mathbf{c})$ so as to maximize $V(\mathbf{l}, \mathbf{c})$. To explicitly solve for the optimal time and consumption allocation problem, it is convenient to define *effective consumption* as $z_t = c_t - r_t$. We redefine the problem as one of finding the optimal values of ℓ_t and z_t in the usual form of a discounted utility model. The next step is to express the budget constraint, (14.5), in terms of z_t. To do so, we use the definition of effective consumption and the dynamics of (14.2) and (14.3) to write

$$c_t = z_t + \sigma s_t + (1 - \sigma) a_t, t = 1, \ldots, T, \quad \text{and} \tag{14.6}$$

$$\begin{aligned} a_t &= \alpha c_{t-1} + (1 - \alpha) a_{t-1} \\ &= \alpha z_{t-1} + \alpha \sigma s_{t-1} + (1 - \alpha\sigma) a_{t-1}, \quad t = 2, \ldots, T + 1. \end{aligned} \tag{14.7}$$

One can then recursively calculate the overall lifetime consumption. In the general case where both α and σ are strictly positive, we have

$$\sum_{t=1}^{T} c_t = \sum_{t=1}^{T} \kappa_t (z_t + \sigma s_t) + \frac{(\kappa_0 - 1)}{\alpha} a_1, \quad \text{where} \tag{14.8}$$

$$\kappa_t = \frac{1 - (1 - \sigma)(1 - \alpha\sigma)^{T-t}}{\sigma}, \quad t = 0, \ldots, T. \tag{14.9}$$

To see this, let C, Z, S, and A denote the summation from $t = 1$ to T of c_t, z_t, s_t, and a_t, respectively. Adding expression (14.6) from 1 to T and expression (14.7)

from 2 to $T+1$ (defining a_{T+1} in the obvious way) yields

$$C = Z + \sigma S + (1-\sigma)A, \quad \text{and}$$
$$A + a_{T+1} - a_1 = \alpha Z + \alpha\sigma S + (1-\alpha\sigma)A.$$

From the second equation, we have that $A = Z/\sigma + S + (a_1 - a_{T+1})/\alpha\sigma$, which we plug into the first equation to obtain

$$C = \frac{1}{\sigma}(Z + \sigma S) + \frac{1-\sigma}{\alpha\sigma}(a_1 - a_{T+1}). \tag{14.10}$$

Using (14.7), one can verify that

$$a_{T+1} = \alpha \sum_{t=1}^{T} (1-\alpha\sigma)^{T-t}(z_t + \sigma s_t) + (1-\alpha\sigma)^T a_1.$$

Replacing a_{T+1} in (14.10) produces (14.8) and (14.9).

If $\sigma = 0$, then we notice that $c_t = z_t + a_t$ and that $a_t = \alpha z_{t-1} + a_{t-1}$. Using induction it follows that

$$\sum_{t=1}^{T} c_t = \sum_{t=1}^{T} (1 + (T-t)\alpha) z_t + T a_1. \tag{14.11}$$

Finally, if $\alpha = 0$, adding expression (14.6) from 1 to T produces

$$\sum_{t=1}^{T} c_t = \sum_{t=1}^{T} (z_t + \sigma s_t) + (1-\sigma) T a_1. \tag{14.12}$$

We assume the general case in which $\alpha, \sigma > 0$. Replacing (14.8) in the left-hand side of (14.5), using $\sum_{t=1}^{T} w_t = T - \sum_{t=1}^{T} \ell_t$ in the right-hand side of (14.5) and rearranging terms produces

$$\max_{(\mathbf{l},\mathbf{z})} \; V(\mathbf{l},\mathbf{z}) = \sum_{t=1}^{T} u(\ell_t) + \sum_{t=1}^{T} v(z_t), \tag{14.13}$$

$$\text{s.t.} \quad \mu \sum_{t=1}^{T} \ell_t + \sum_{t=1}^{T} \kappa_t z_t \le \mu T - \sum_{t=1}^{T} \sigma \kappa_t s_t - \frac{\kappa_0 - 1}{\alpha} a_1. \tag{14.14}$$

The first order conditions are

$$u'(\ell_t) = \mu\lambda, \;\; t = 1, \ldots, T, \text{ and} \tag{14.15}$$
$$v'(z_t) = \kappa_t \lambda, \;\; t = 1, \ldots, T. \tag{14.16}$$

It is interesting to examine expression (14.14). The left-hand side contains the drivers of utility: leisure time and effective consumption. The wage rate increases not only the price of leisure (in reality, it makes consumption more affordable) but also the maximum budget, μT. Effective consumption is multiplied by the

coefficient, κ_t, which is easy to see from (14.9) that it is decreasing in t. If we interpret this coefficient as a price, we observe that effective consumption is more expensive to purchase at the beginning of the planning horizon than at the end. The reason for this, of course, is that early consumption above the adaptation level increases future adaptation levels.

The right-hand side of (14.14) contains the constraints of the drivers of utility. The main constraint is the total money that could be earned if all available time were to be spent working, μT. This maximum budget is reduced by (a weighted sum of) the social comparison level and the initial adaptation level. Subsequent adaptation levels are not included, as they follow endogenously from the optimization program. In summary, *social comparison and current adaptation reduce the available budget*.

We assume that the right-hand side of the modified budget constraint (14.14) is non-negative. It follows from (14.15) that the optimal time allocated to leisure, ℓ_t, is the same in every period. Let ℓ denote this constant value. The remaining time is devoted to work, $w = 1 - \ell$, which is also constant.

We now examine (14.16). Knowing that κ_t is decreasing and that v' is strictly decreasing implies that the optimal effective consumption, z_t, is necessarily increasing over time. To ensure that $z_1 \geq 0$, it is sufficient to have $v'(0^-) \geq \kappa_1 u'(0)/\mu$. That effective consumption is increasing is intuitive. Recall that consumption above the adaptation level yields positive utility during the current period, but lowers utility during the subsequent periods as it increases the adaptation levels. This negative effect fades the closer one gets to the final period. Hence, optimal planning induces increasing values of z_t. Of course, increases in z_t produce increases in c_t, as is evident from expression (14.8). This expression shows that an increase in z_t directly translates to an increase in c_t and an additional increase in $c_{t+1}, \ldots, c_T$. Hence, consumption increases more than effective consumption.

In the optimal plan, a decision maker follows a regular schedule of w hours of work and ℓ hours of leisure. Both consumption and effective consumption are increasing, which means saving in early periods, followed by borrowing later in life. If the consumption good is not adaptive, $\alpha = 0$, and there is no social comparison, $\sigma = 0$, then it follows from (14.6) that consumption and effective consumption are constant, as $c_t = z_t + a_1$.

It is possible to find a closed form solution if both u and v take a power form with the same exponent β, that is, $u(\ell) = \ell^\beta$ and $v(z) = z^\beta$, $\ell, z \geq 0$. In this case,

$$\ell = \frac{\mu T - \sum_{t=1}^{T} \sigma \kappa_t s_t - ((\kappa_0 - 1)/\alpha) a_1}{\mu T + \mu^{1/(1-\beta)} \sum_{t=1}^{T} (1/\kappa_t)^{\beta/(1-\beta)}} \quad \text{and} \tag{14.17}$$

$$z_t = \frac{\mu T - \sum_{t=1}^{T} \sigma \kappa_t s_t - ((\kappa_0 - 1)/\alpha) a_1}{\kappa_t^{1/(1-\beta)} (1/\mu)^{\beta/(1-\beta)} T + \kappa_t^{1/(1-\beta)} \sum_{t=1}^{T} (1/\kappa_t)^{\beta/(1-\beta)}}. \tag{14.18}$$

Assuming $\beta > 0$, we verify that time spent on leisure decreases with social comparison level, initial adaptation, and wage. In contrast, effective consumption increases with wage. Actual consumption can be derived from effective consumption using (14.6) and (14.7).

14.3 Income–Happiness Relationship

Total utility in our model is regarded as an empirical approximation of happiness. Aristotle believed that happiness must be judged over a lifetime and that its constituent parts included wealth, relationships, and bodily excellences (e.g., health and beauty). To Bentham [5], happiness was attained by maximizing the positive balance of pleasure over pain as measured by experienced utility [29]. He argued that human affairs should be arranged to attain the greatest happiness for the greatest number of people.

In recent years, researchers have been able to measure happiness and have collected a great deal of empirical data that relates income, as well as other social and biological factors, to happiness. Happiness in these surveys is measured by *asking* people how satisfied they are with their lives. A typical example is the *General Social Survey* [12], which asks "Taken all together, how would you say things are these days—Would you say that you are very happy, pretty happy, or not too happy?" In the World Values Survey, Inglehart and colleagues [24] use a 10-point scale with 1 representing dissatisfied and 10 representing satisfied to measure well-being. Pavot and Diener [41] use five questions each rated on a scale from one to seven to measure life satisfaction.

Davidson et al. [9, 11] have found that when people are cheerful and experience positive feelings (e.g., funny film clips), there is more activity in the front left section of their brains. The difference in activity between the left and right sides of the prefrontal cortex seems to be a good measure of happiness. Self-reported measurements of happiness correlate with this measure of brain activity, as well as with ratings of one's happiness made by friends and family members [33]. Diener and Tov [13] report that subjective measures of well-being correlate with other types of measurements of happiness, such as biological measurements, informant reports, reaction times, open-ended interviews, smiling behavior, and online sampling. Kahneman et al. [26] discuss biases in measuring well-being that are induced by using a focusing illusion in which the importance of a specific factor (e.g., income, marriage, health) is exaggerated by drawing attention to it. Nevertheless, Kahneman and Krueger [25] argue that self-reported measures of well-being may be relevant to future decisions, as idiosyncratic effects are likely to average out in representative population samples. Frey and Stutzer [21] conclude as follows: "The existing research suggests that, for many purposes, happiness or reported subjective well-being is a satisfactory empirical approximation to individual utility."

If people pursue the goal of maximization of happiness and have reported their happiness levels truthfully in the variety of surveys discussed above, then how do we explain that happiness scores have remained flat in spite of significant increases in real income over time (Figure 14.1)? Of course, happiness depends on factors other than income such as the genetic makeup of a person, family relationships, community and friends, health, work environment (unemployed, job security), external environment (freedom, wars or turmoil in society, crime), and personal values (perspective on life, religion, spirituality). Income, however, does influence an individual's happiness up to a point and has a moderating effect on the adverse effects of

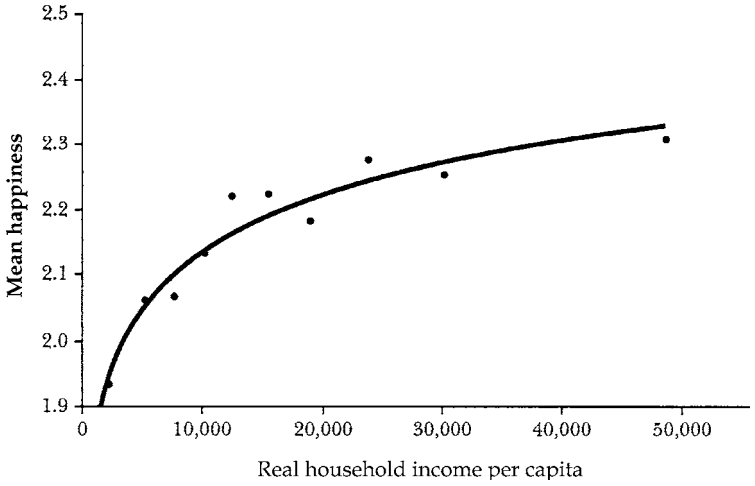


Fig. 14.3 Mean happiness and real household income for a cross-section of Americans in 1994. *Source:* diTella and MacCulloch [14]

some life events [52]. As shown in Figure 14.3, mean happiness for a cross-section of Americans does increase with income, though at a diminishing rate. In fact, richer people are substantially happier relative to poorer people in any given society.

Our time allocation model is consistent with the joint empirical finding that happiness over time does not increase appreciably in spite of large increases in real income, but happiness in a cross-section of data does depend on relative levels of income. That rich people are happier than poor people at a given time and place is easy to justify even by the Discounted Utility Model. Income effects are magnified if the reference level depends on social comparison as, by and large, richer people have a favorable evaluation of their own situation compared to others. Over time, though, both rich and poor people have significantly improved their living standards, but neither group has become happier. Adaptation explains this paradoxical finding.

Consider Mr. Yoshi, a young professional living in Japan in the 1950s. He was content to live in his parents' house, drive a used motorcycle for transportation, wash his clothes in a sink and listen to the radio for entertainment. Also consider Ms. Yuki, a young professional living in Japan in the 1990s. She earns five times the income of Mr. Yoshi in real terms. She wants her own house, automobile, washing machine, refrigerator, and television. She travels abroad for vacation and enjoys expensive international restaurants. Because Mr. Yoshi and Ms. Yuki are in similar social positions for their times, then both will have the same level of happiness. Happiness does not depend on the absolute level of consumption, which is substantially higher for Ms. Yuki. Instead, happiness depends on the level of consumption relative to the adaptation level. Ms. Yuki has become adapted to a much higher level of consumption and therefore finds that she is no happier than Mr. Yoshi. In our time allocation model, as the wage rate (μ) increases, total utility stands still if the

initial reference point (r_1) also increases in the same manner calculated by the model. Thus, the "Easterlin Paradox"—that happiness scores have remained flat in developed countries despite considerable increases in average income—can be explained by the total utility maximization, provided the initial reference level, which measures expectations, increases with prosperity. Happiness scores for poorer countries have in fact increased over time as the increased income has provided for additional basic goods such as adequate food, shelter, clean water, and health care.

Many authors have given a qualitative argument that the reference point is higher for a person living in 1990s Japan than in 1950s Japan. Actually, we now show that as μ increases, total utility stands still if a_1 increases. In the following numerical example, we set $\alpha = 1$ and $\sigma = 0$. An individual with $a_1 = 0$ and $\mu = 1$ would obtain a total optimal utility of 11.4. This is obtained by solving the leisure–consumption problem (14.1) assuming the power form for u and v with exponent 0.5. This same optimal total utility is obtained by setting $\mu = 5$ and $a_1 = 3.4$. Thus, a substantial increase in wage does not lead to an increase in total utility if the initial reference level has also increased.

So far, we have seen that our time allocation model is consistent with empirical findings that within a country richer people are happier than poorer people, but, for prosperous countries, well-being does not increase over time in spite of permanent increases in income for all. In a survey in the United States, when asked to specify a single factor that would most improve their quality of life, the most frequent answer was "more money." Thus, the puzzle remains: why do people believe more money will buy them more happiness when in fact it may not. There is also some evidence that people are working harder at the expense of leisure; sleep time has gone down from 9.1 h per night to 6.9 h per night during the 20th century. The misallocation of time between work and leisure is difficult to prove, but we will show that under the plausible psychological assumption of projection bias such a misallocation is indeed possible.

14.4 Predicted Versus Actual Happiness

> The great source of both the misery and disorders of human life, seems to arise from overrating the difference between one permanent situation and another.
>
> — Adam Smith (1759, Part III, Chapter III]

If people plan optimally, then they will maximize happiness by appropriately balancing time devoted to work and to leisure and by choosing an increasing consumption path. Optimal planning, however, requires that one correctly predict the impact of current consumption on future utility. An increase in consumption has two perilous effects on future utility. First, the adaptation level goes up and therefore future experienced utility declines (e.g., people get used to a fancier car, a bigger house, or vacation abroad). Second, the social comparison level may go up, which again reduces experienced utility. When one joins a country club or moves to a more prosperous neighborhood, the peer group with which social comparisons are made

changes. The individual now compares himself with more prosperous "Joneses" and comparisons to his previous peer group of less prosperous "Smiths" fades. If the individual foresees all this, then he can appropriately plan consumption over time and realize higher total utility in spite of a higher level of adaptation and an upward movement in peer group. The rub is that people underestimate adaptation and changes in peer group. Loewenstein et al. [35] have documented and analyzed underestimation of adaptation and have called it *projection bias*.

Because of projection bias, an individual will realize less happiness than predicted. The gap between predicted and actual levels of happiness (total utility) further increases if one plans myopically rather than optimally. An example of a myopic plan is to allocate a budget or income equally in each period (constant consumption), as opposed to an increasing plan. A worse form of myopic planning would be to maximize immediate happiness through splurging (large consumption early on) which is what some lottery winners presumably end up doing.

We buy too much when hungry [40], forget to carry warm clothing during hot days for cooler evenings, predict that living in California will make us happy [48], and generally project too much of our current state into the future and underestimate adaptation [22, 34, 36]. vanPraag and Frijters [57] estimate a rise of between 35 and 60 cents in what one considers required income for every dollar increase in actual income. Stutzer [54] also estimates an increase in adaptation level of at least 40 cents for each dollar increase in income. After the very first year, the joy of a one-dollar increase in income is reduced by 40%, but people are unlikely to foresee this reduced contribution to happiness. People do qualitatively understand that some adaptation to the change in lifestyle that comes with higher income will take place; they simply underestimate the magnitude of the changes.

In our model, the chosen consumption plan determines the actual reference level, r_t, by means of (14.2) and (14.3). In every period, an individual observes the current reference level, but may fail to correctly predict the value of this state variable in future periods. According to projection bias, the predicted reference level is somewhere between the current reference level and the actual reference level. The relationship between the actual and predicted reference levels can be modeled using a single parameter, π, as follows:

$$\begin{aligned}\text{Predicted reference level} = {} & \pi(\text{current reference level}) \\ & + (1-\pi)(\text{actual reference level}).\end{aligned}$$

Thus, when $\pi = 0$, there is no projection bias, and the predicted reference level coincides with the actual reference level. If $\pi = 1$, then the individual adopts the current reference level as the future reference level. An intermediate value of $\pi = 0.5$ implies that the individual's predicted reference level is halfway between the current and actual reference levels. This projection bias model can be extended to any state variables that influence preferences, such as satiation level [3]. If consumption stays above the actual reference level over time, then an individual with projection bias may be surprised that the actual, realized utility in a future period is lower than what was predicted. The reason, of course, is that the actual reference level is higher than

anticipated. Actual happiness associated with higher levels of consumption may be much lower than what was hoped for. This gap may motivate an individual to work even harder to increase income in the hopes of improving happiness. But this chase for happiness through higher and higher consumption is futile if the reference level keeps increasing.

To formalize these ideas, let τ be the current period. The actual and predicted reference levels for a subsequent period t are r_t and $\hat{r}_{\tau,t}$, respectively. Now,

$$\hat{r}_{\tau,t} = \pi r_\tau + (1-\pi) r_t,$$

for which r_t follows the dynamics governed by (14.2) and (14.3). The *actual* utility is given by the chosen consumption plan according to the time allocation model; however, the chosen consumption plan might not be the optimal one. The reason for this is that during period τ, the individual will maximize the *predicted* utility given by

$$\hat{V}_\tau(\ell_\tau, \ell_{\tau+1}, \ldots, \ell_T; c_\tau, c_{\tau+1}, \ldots, c_T | r_\tau, \pi) = \sum_{t=\tau}^{T} u(\ell_t) + \sum_{t=\tau}^{T} v(c_t - \hat{r}_{\tau,t}). \quad (14.19)$$

The difference between the actual and the predicted utility can be demonstrated by a simple example. Figure 14.4 compares the optimal plan to the plan implemented by an individual experiencing the most extreme form of projection bias, namely, $\pi = 1$ and $\alpha = 1$. In this example, wage is set to one, and both $u(x)$ and $v(x)$ are set to $\sqrt{x}$.

The optimal consumption plan exhibits an accelerating, increasing pattern, as argued in Section 14.2. This is indeed rational for an individual who is fully aware of two facts: (1) increments and not absolute levels are the drivers of utility of consumption and (2) high consumption at the beginning of the time horizon heavily taxes utility in later periods, as it raises the adaptation level in a permanent way. Hence, it is no surprise that consumption is low in the beginning and high toward the end of the planning horizon. As expected, the optimal time for work and leisure is constant over time.

A rational individual would allocate approximately 80% of his time to leisure and 20% to work. Now, consider the projection bias plan; the consumption plan under projection bias begins in period 1 with a plan to consume 0.5 units. The amount of time devoted to work and leisure is the same, i.e., 50% to work and 50% to leisure. This is not a coincidence. If $\pi = 1$, then the individual predicts that the reference point for consumption will remain constant; therefore, this individual treats both leisure and consumption as basic goods. As u and v are identical, an equal allocation of time to work and leisure is optimal. Moreover, the individual plans to maintain the constant level of consumption of five units per period.

In period 2, the individual realizes that the reference level, r_2, is higher than r_1; in fact, $r_2 = c_1 = 0.5$. This is a cause of concern, as the original plan of flat consumption of 0.5 units will yield zero utility, $v(0.5-0.5) = 0$ for the consumption component. Here, projection bias enters again. The individual again predicts that the

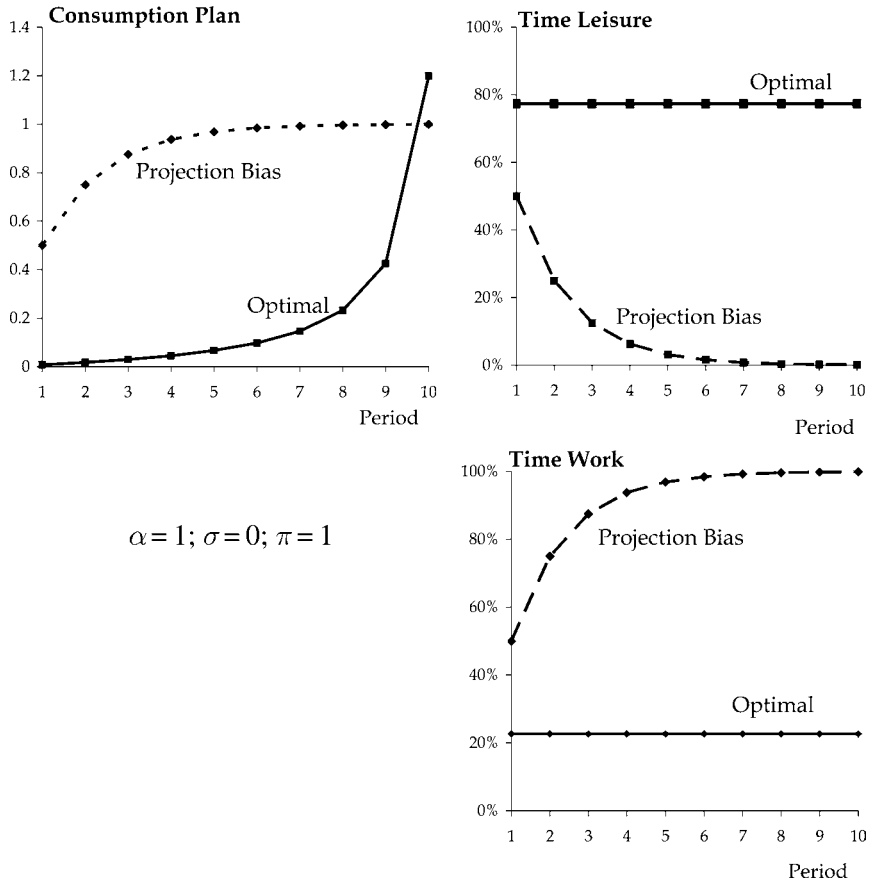


Fig. 14.4 Impact of projection bias on time allocation [$\alpha = 1, \pi = 1, \mu = 1$]

future reference level will be the same as the current reference level of 0.5 units. The individual, therefore, hopes that by increasing consumption above 0.5 units, he can obtain higher utility. But to do so, he needs to expand the budget, which is not a problem because he can work for 0.75 units, instead of 0.5. The additional units of time are taken from leisure time, which now decreases to 0.25 units. In period 3, the same process repeats itself. The gap between the actual and the predicted reference level may motivate the person to work even harder to increase income in the hopes of improving happiness. But this chase for happiness through higher and higher consumption is futile as the reference level keeps on increasing. Actual happiness associated with higher levels of consumption may be much lower than what was hoped for.

The degree of misallocation of time between work and leisure depends on both the adaptation factor, α, and the projection bias parameter, π. In our example, percentage time allocations to work for various combinations of α and π are shown

Table 14.1 Percent of time allocated to work [$\mu = 1$]

	Optimal	Projection Bias		
Adaptation Factor	$\pi = 0$	$\pi = 0.1$	$\pi = 0.5$	$\pi = 1.0$
$\alpha = 0.1$	42	43	50	60
$\alpha = 0.5$	28	32	54	81
$\alpha = 1.0$	23	28	64	90

in Table 14.1. For the optimal plan, as the adaptation rate increases, the percentage of time allocated to work decreases. Similarly, for a given α, as projection bias increases, the individual works harder. In all cases, the actual total utility under projection bias will be lower than that given by the optimal plan because of the misallocation of time and the excessive consumption in early periods.

14.5 Higher Pay—Less Satisfaction

So far we have demonstrated that projection bias could induce people to work harder and therefore be left with less leisure time compared to the rational plan. We now examine the effects of increases in wage rate on total utility. A rational individual will always experience a higher total utility with a higher wage rate by judiciously allocating time between work and leisure. Individuals, however, do not always make sensible tradeoffs between work and leisure. Average sleep hours in the United States fell from 9 h per night in 1910 to 7.5 h per night in 1975 with a further decline to 6.9 h per night between 1975 and 2002. A *USA Today* report on May 4, 2007 titled "U.S. Workers Feel Burn of Long Hours, Less Leisure" reports that US workers put in an average of 1,815 h in 2002 compared to European workers who ranged from 1,300 to 1,800 h (see also [32, p. 50]). Schor [49] argues that Americans are overworked. In some professions in which the relationship between income and hours worked is transparent (e.g., billable hours for lawyers and consultant), there is a tendency to allocate relatively more time to work due to peer pressure.

A theory in anthropology holds that the rise of civilization is the consequence of the increased availability of leisure time [23]; Sahlins [46, pp. 85–89] argues that the quantity of leisure time proxies for well-being. Putnam [43] observed in his book, *Bowling Alone*, that people who engage in leisurely activities with others were, on average, happier than those who spent their leisure time alone. Aguiar and Hurst [2], who document an increase in leisure time for less educated people, observe that there has been a substantial increase in time spent watching television (passive leisure) and a significant decline in socializing (active leisure) for people of all education levels from 1965 to 2003.

It is possible that experienced utility in a given period $u_t + v_t$ may be lower if one disproportionately allocates more time to work at the expense of leisure. Budding entrepreneurs, investment bankers, and executives of technology companies may complain about their "all work and no play" lifestyle, but many of them do retire

early or change careers and it is hard to argue that their excessive work in the early part of their careers was not rational. All work and no play may make Jack a dull boy, but if that is what Jack desires then there can be no disputing his taste. We show that in the presence of projection bias, an individual may reduce his actual *total* utility by choosing a higher wage option. A simple, two-period example will suffice to illustrate this paradoxical result.

Consider a two-period example with $\alpha = 1$ and $\pi = 1$. In period 1, an individual maximizes predicted utility over the two periods by planning to work $w_{1,1}$ in period 1 and $w_{1,2}$ in period 2. Because leisure is a basic good, the individual plans an equal amount of leisure in each period. Consequently, the amount of work in each period is also equal, i.e., $w_{1,1} = w_{1,2}$. Under extreme projection bias, $\pi = 1$, the individual considers that consumption also behaves as a basic good. Hence, the per-period consumption corresponds to the budget generated for that period, namely, $\mu w_{1,1}$. Finally, $w_{1,1}$ is found by optimizing the predicted total utility given by

$$V(\ell, w) = 2[u(1 - w_{1,1}) + v(\mu w_{1,1})]. \tag{14.20}$$

The first-order condition is given by

$$u'(1 - w_{1,1}) = \mu v'(\mu w_{1,1}). \tag{14.21}$$

The individual solves this problem and decides on his allocation of budget to leisure and consumption.[2] During the second period, the adaptation level takes the value $r_2 = \mu w_{1,1}$.[3] The individual then realizes that the utility of consumption in period 2 will be zero if he stays with the original plan. He therefore revises the plan by maximizing the utility in period 2:

$$V(w, \ell) = u(1 - w_{2,2}) + v(\mu(w_{2,2} - w_{1,1})). \tag{14.22}$$

The optimal time spent working in period 2, $w_{2,2}$, is the solution to the first-order condition:

$$u'(1 - w_{2,2}) = \mu v'(\mu(w_{2,2} - w_{1,1})). \tag{14.23}$$

Inspecting (14.21) and (14.23), we observe that if $v'(0^+) > u'(1)$, then $w_{1,1}$ is strictly positive *and* $w_{2,2}$ is strictly larger than $w_{1,1}$. Therefore, the individual always revises the plan in favor of increasing work and reducing leisure for the second period. The increase in work in the second period is bounded, as $w_{2,2} - w_{1,1} \leq w_{1,1}$, with strict inequality if u is strictly concave.[4] Thus, the utility from consumption

[2] Applying the implicit function theorem to the first-order condition (14.21), it follows that $w_{1,1}$ increases with μ if and only if the Arrow–Pratt measure of relative risk aversion of v is less than 1. This same condition also applies to $w_{2,2}$, the time that the individual decides to work in period 2 after re-optimizing the predicted utility.

[3] The conclusions and insights are the same if we use the full model and let $r_2 = \sigma s_2 + (1 - \sigma)\alpha\mu w_{1,1}$.

[4] If $w_{2,2} > w_{1,1}$, then using (14.21) and (14.23) yields $\mu v'(\mu w_{1,1}) = u'(1 - w_{2,2}) \geq u'(w_{1,1}) = \mu v'(\mu(w_{2,2} - w_{1,1}))$. As v' is non-increasing, it follows that $w_{2,2} - w_{1,1} \leq w_{1,1}$.

obtained in period 2, in spite of revising the plan, is less than or equal to the predicted utility $v(\mu w_{1,1})$.

The actual total utility is given by

$$u(1-w_{1,1})+v(\mu w_{1,1})+u(1-w_{2,2})+v(\mu(w_{2,2}-w_{1,1})). \tag{14.24}$$

It is clear that the actual total utility (14.24) is lower than the predicted total utility (14.20). In period 1, actual and predicted utilities coincide. However, in period 2, the actual utility of leisure is lower than the predicted utility of leisure ($w_{2,2} > w_{1,1}$). Similarly, in period 2, the actual utility of consumption is lower than the predicted utility of consumption ($w_{2,2} - w_{1,1} < w_{1,1}$). We now show that the misallocation of time between work and leisure could lower actual total utility when the wage rate increases.

In the particular case that u is linear and $v(x) = x^{\beta}$, $x \geq 0$, the actual utility is *increasing* in μ if $\beta < 2/3$ and is *decreasing* in μ if $\beta > 2/3$. That actual utility may be decreasing with wage rate is puzzling. To see this, notice that planned work is given by

$$w_{1,1} = \mu^{\beta/(1-\beta)}\beta^{1/(1-\beta)} \quad \text{and} \quad w_{2,2} = 2w_{1,1},$$

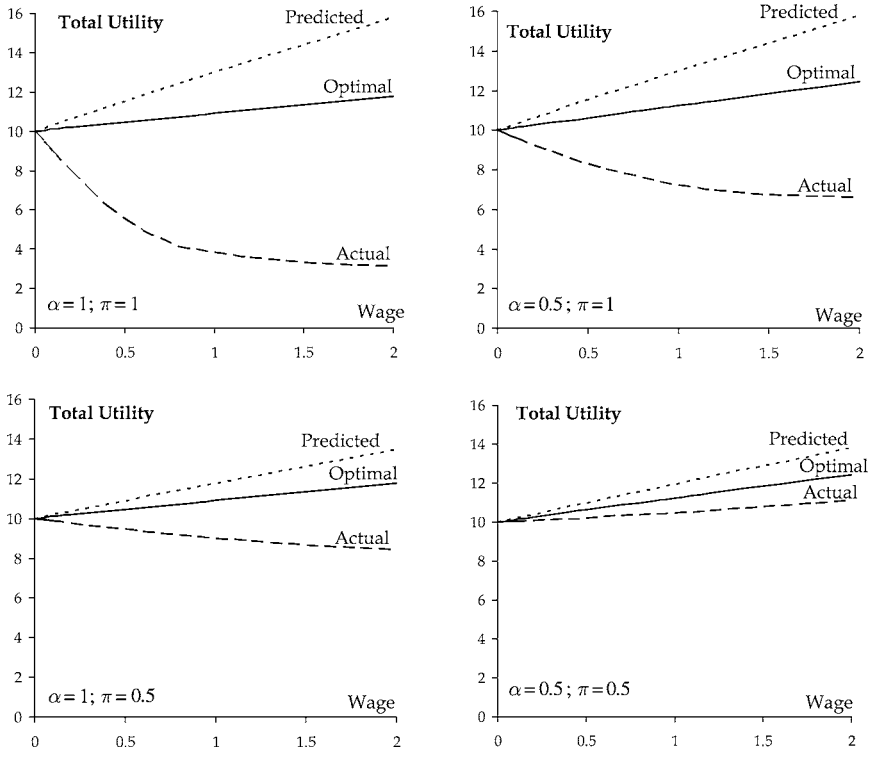


Fig. 14.5 Impact of wage rate on total utility under projection bias [$T = 10$, $u(\ell) = \ell^{0.8}$, $v(z) = z^{0.5}$, $\sigma = 0$]

which, when plugged into the equation for actual utility, yields

$$2+(2-3\beta)(\mu\beta)^{\beta/(1-\beta)}. \tag{14.25}$$

The puzzling result that total utility can be decreasing with wage rate holds more generally. Figure 14.5 shows the relationship between total utility and wage rate for a 10-period case ($T = 10$) in which both u and v are strictly concave (taking power forms with exponents 0.8 and 0.5, respectively). Optimal total utility is, of course, always increasing with wage rate, but projection bias may decrease the actual total utility as shown in the upper left panel of Figure 14.5.

One must therefore be deliberate in choosing a high wage career (e.g., consulting or investment banking) and be mindful of Veblen's [58] observation: "But as fast as a person makes new acquisitions, and becomes accustomed to the resulting new standard of wealth, the new standard forthwith ceases to afford appreciably greater satisfaction than the earlier standard did."

14.6 Social Comparison

Adam Smith [50] stated "With the greater part of rich people, the chief enjoyment of riches consists in the parade of riches." Veblen [58] echoes a similar sentiment: "The tendency in any case is constantly to make the present pecuniary standard the point of departure for a fresh increase of wealth; and this in turn gives rise to a new standard of sufficiency and a new pecuniary classification of one's self as compared with one's neighbors." Meaning, because most rich people pursue comparative ends, they will ultimately fail to become happier.

An immediate question arises whether one can improve one's happiness simply by imagining less fortunate people. However, Kahneman and Miller [27] assert that to influence our hedonic state, counterfactuals must be plausible, not just possible, alternatives to reality. The all too common tactic of a parent coaxing a child to appreciate food by reminding them of starving children in third world countries does not work. There seems to be a tendency to want conspicuous success. In many professions, income has become that measure of success; therefore, people pursue higher income not just for consumption, but as a scorecard of their progress. Conspicuous success also seems to have no end. Russell [44] wrote, "If you desire glory, you may envy Napoleon. But Napoleon envied Caesar, Caesar envied Alexander, and Alexander, I dare say, envied Hercules, who never existed."

Social comparison levels in our model are exogenous, though a theory in which the appropriate peer group and social comparison level is endogenous would be useful. Nevertheless, we can provide some insight into the influence of social comparison on happiness. Consider, for example, three groups of people: those in the highest quintile, in the lowest quintile, and at the median level of income ($83,500, $17,970, and $42,228, respectively, for the United States in 2001). By and large,

richer people have a favorable evaluation of their own situation compared to others. In contrast, the economically disadvantaged will have an unfavorable evaluation of their relative position in society. Assume that the social comparison level, S, is equal to the median income. For simplicity, we assume constant consumption around the annual income for each group. If we focus only on the utility of consumption, then without social comparison ($\sigma = 0$) each of the three groups will converge to the neutral level of happiness as each becomes adapted to their own past consumption levels. By including social comparison, the happiness levels are pulled toward, but do not converge on, the neutral level. The long run experienced utility is given by $v(\sigma(x-m))$, which is the median income. This heuristic argument is consistent with the empirical finding that richer people are happier than poorer people.

Now consider two individuals: Average Joe and Fantastic Sam. Average Joe is a highly paid stockbroker ($\mu = 10$), but his peer group also has high incomes ($S = 8$). Assume that $u(x) = v(x) = \sqrt{x}$, $\alpha = 1$, $\sigma = 0.5$, and $a_1 = 0$. In an optimal plan, Average Joe would devote 96% of his available time to work and 4% to leisure. His total consumption would be 96 units and his total utility would be 13.8. In contrast, Fantastic Sam is an above average journalist who earns half as much as Joe ($\mu = 5$), but compares favorably with his peer group ($S = 1$). Planning optimally, Sam would devote 80% of his time to work and 20% to leisure. His total consumption would be 40 units and his total utility would be 17.89. Sam would be happier than Joe in spite of his lower income and lower consumption because his position relative to his peers is superior to that of Joe's.

Projection bias could induce Sam to chase the prosperous life of a stockbroker if offered the opportunity. In this case, projection bias would affect him through his underestimation of the upcoming change in social comparison level. Sam could indeed be happier as a stockbroker, but he should put some thought into forecasting his relative position amongst stockbrokers and how that would impact his future utility. If he concludes that he would be an average stockbroker, then journalism might indeed be the right pond for Fantastic Sam [17].

14.7 Reframing

> One does not become happy overnight, but with patient labor day after day. Happiness is constructed, and that requires effort and time. In order to become happy, we have to learn how to change ourselves.
>
> — Luca and Francesco Cavalli-Sforza (1998)

In our model, the dynamics of adaptation and social comparison are not part of an individual's choices. This implies that an individual does not have control over adaptation to consumption or over one's own expectations determined by his peer group. It is possible to have heterogeneous individuals with different speeds of adaptation and weights given to social comparison. However, for a given individual, both α and

σ are fixed, and there is nothing this individual can do to change his speed of adaptation or intensity of social comparison. The same can be said about π, the inability to accurately predict future reference levels.

While adaptation and social comparison are unavoidable to a certain extent, we believe that individuals do have some tools available to moderate these factors. It is possible that through reframing activities such as spiritual practices, meditation, or prayer, one might gain a better perspective on life and reduce the harmful effects of comparison. Such practices, however, require considerable time, effort, and discipline. An admiring fan congratulated a violinist for playing so beautifully and said "I would love to play like you." The violinist answered: "Yes, but would you love it even if you had to practice 10,000 h?"

We now attempt to introduce the impact of reframing and perspective seeking into our model. We assume that a new decision variable is available to the individual, namely the time that he sets aside in each period for "reframing activities." To keep things simple, we assume that this time is constant throughout the planning horizon, which we denote by q.

The choice of q is made in period 1, and after this choice is made the time available for work and leisure is reduced to $1-q$ in all periods. In other words, an individual commits in period 1 to set aside a fixed amount of time to such practices. Reframing activities contribute to gaining perspective on life, appreciating all received goods as if had been received for the first time, encountering ways to suppress or avoid (unfavorable) social comparison and finding inner happiness. Lama and Cutler [31] explain "The actual secrets of the path to happiness are determination, effort, and time." Neuroscience confirms that repetition is essential for the brain to be retrained. Cellists have more developed brain areas for the fingers of their left hand, mechanics for their sense of touch, and monks for the activity in the left prefrontal cortex, which is associated with cheerfulness.

Devoting time to reframing activities has an opportunity cost (less time available for work or leisure). We assume that the benefit of reframing activities is in lowering the reference level. Specifically, we modify the time allocation model by replacing and updating (14.2) with

$$r_t = e^{-\rho q}[\sigma s_t + (1-\sigma)a_t], \quad t = 1, \ldots, T,$$

where ρ measures the effectiveness of reframing activities (e.g., competent teacher, seriousness of commitment) and q is the time devoted to such activities. The modification simply multiplies the previous reference level by a reduction factor, $e^{-\rho q}$. This reduction factor is 1 if the time spent in reframing activities is 0; however, if $q > 0$, then the factor is strictly less than 1. The value of q is now part of the set of decision variables.

It is possible that unless ρ is larger than a certain threshold value, the individual may find that it is not worth spending any time in reframing activities. This is illustrated in Figure 14.6. Note that the optimal time spent in reframing activities is non-monotonic with ρ. This is to be expected. If ρ is sufficiently high, then a little time devoted to reframing can do a lot to reduce reference levels. Of course, total

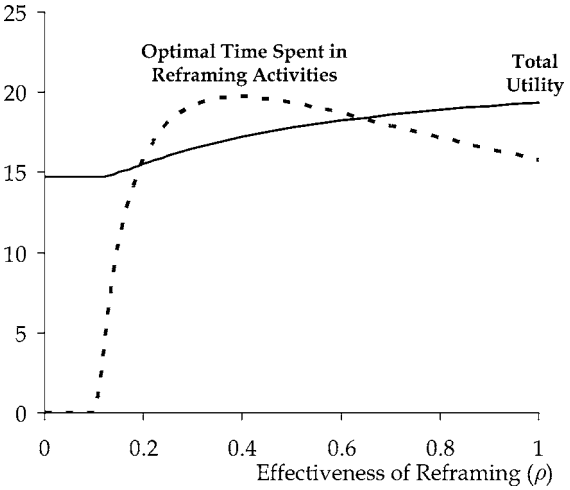


Fig. 14.6 Total optimal time spent on spiritual practices and total utility as a function of the effectiveness of these practices [$S = 5$, $\sigma = 0.5$, $\alpha = 1$]

utility is monotonic with ρ, as the per-period utility of consumption increases as reference levels decrease.

14.8 Conclusions

> No society can surely be flourishing and happy, of which the far greater part of the members are poor and miserable.
>
> — Adam Smith (1776)

A rational individual chooses an appropriate trade-off between work and leisure, thereby maximizing happiness. In this chapter, we have proposed a simple adaptation and social comparison model of time allocation, which predicts that happiness increases with income at a diminishing rate. Furthermore, the optimal consumption path is increasing over time, as is relative consumption over the reference level.

Our model is consistent with the empirical findings that richer people are happier than poorer people, but that happiness scores have remained flat over time in spite of astonishing increases in real income. Perhaps, the most interesting implications of our model are obtained under the assumption that people underestimate the rise in their reference level (due to projection bias) and thus overestimate the utility of consumption. Projection bias may lead an individual to devote too much time to work at the expense of leisure. Their predicted utility under projection bias is higher than the actual realized utility. This is why we believe that more money will buy us more happiness when in fact it may not. Because of their misallocation of time

between work and leisure, the actual realized utility may even decline at higher wage rates.

In a preliminary attempt, we show that reframing activities, such as meditation or other spiritual practices, may improve happiness, but these activities require a commitment of time. Davidson and Harrington [10] find that the happiness level of Buddhist monks is higher than the average population in spite of their frugal lifestyle. Additional empirical and theoretical work is needed to understand the influence of reframing activities on moderation of reference levels.

Projection bias diverts resources from leisure toward adaptive consumption. Great discipline is therefore required to give adequate attention to the importance of leisure (e.g., time spent with family and friends, sleep, and exercise). We are reluctant to venture into policy prescriptions without a thorough analysis. However, if there is no awareness of projection bias, then a judicious application of policies like mandatory leave (2 weeks in the United States versus 6 weeks in France), restrictions on work hours within limits (recent reforms for medical residents), having higher sales taxes for adaptive goods than for basic goods, and family friendly practices, such as flexible hours, could improve happiness. Time is the ultimate finite resource; therefore, its allocation between work and leisure to improve happiness needs further empirical and theoretical inquiry. Restoring a harmonious balance between work and leisure is a precondition to "catching" the elusive goal of happiness.

References

1. Abdellaoui M, Bleichrodt H, Paraschiv C (2007) Loss aversion under prospect theory: A parameter-free measurement. Management Science 53:1659–1674
2. Aguiar M, Hurst E (2006) Measuring trends in leisure: The allocation of time over five decades. NBER Working paper n.12082
3. Baucells M, Sarin R (2006) Predicting utility under satiation and habituation. IESE Business School. http://webprofesores.iese.edu/mbaucells/
4. Baucells M, Sarin R (2007) Satiation in discounted utility. Operations Research 55(1): 170–181
5. Bentham J (1789) Principles of morals and legislation. Clarendon Press, Oxford
6. Brickman P, Coates D, Janoff-Bullman R (1978) Lottery winners and accident victims: Is happiness relative? Journal of Personality and Social Psychology 36:917–927
7. Cavalli-Sforza L, Cavalli-Sforza F (1998) La science du bonheur. Odile Jacob, Paris
8. Clark AE (1996) Job satisfaction in Britain. British Journal of Industrial Relations 34: 189–217
9. Davidson R, Jackson D, Kalin N (2000) Emotion, plasticity, context, and regulation: Perspectives from affective neuroscience. Psychological Bulletin 126:890–906
10. Davidson RJ, Harrington A (2001) Visions of compassion: Western scientists and Tibetan Buddhists examine human nature. Oxford University Press, Oxford
11. Davidson RJ, Kabat-Zinn J, Schumacher J, Rosenkranz M, Muller D, Santorelli SF, Urbanowski F, Harrington A, Bonus K, Sheridan JF (2003) Alterations in brain and immune function produced by mindfulness meditation. Psychosomatic Medicine 65:564–570

12. Davis JA, Smith TW, Marsden PV (2001) General social surveys, 1972–2000, Cumulative codebook. Roper Center for Public Opinion Research, Storrs, CT
13. Diener E, Tov W (2005) National subjective well-being indices: An assessment. In: Land KC (ed) Encyclopedia of social indicators and quality-of-life studies. Springer, New York
14. diTella R, MacCulloch R (2006) Some uses of happiness data in economics. Journal of Economic Perspective 20(1):25–46
15. Easterlin RA (2003) Explaining happiness. Proceedings National Academy Sciences 100(19):11176–11186
16. Easterlin RA (1995) Will raising the incomes of all increase the happiness of all? Journal of Economic Behavior Organization 27:35–48
17. Frank RH (1985) Choosing the right pond. Oxford University Press, New York
18. Frank RH (1997) The frame of reference as a public good. The Economic Journal 107(445):1832–1847
19. Frank (1999) Luxury fever. Princeton University Press, Princeton
20. Frederick S, Loewenstein G (1999) Hedonic adaptation. In; Kahneman D, Diener E, Schwarz N (eds) Well being: The foundation of hedonic psychology. Russell Sage, New York, 302–329
21. Frey BS, Stutzer A (2002) What can economists learn from happiness research. Journal of Economic Literature 40(2):402–435
22. Gilbert D (2006) Stumbling on happiness. Alfred A. Knopf, New York
23. Gross DR (1984) Time allocation: A tool for the study of cultural behavior. Annual Review of Anthropology 13:519–559
24. Inglehart R, et al (2000) World values surveys and European values surveys, 1981–84, 1990–93, 1995–97. Institute for Social Research, Ann Arbor, MI
25. Kahneman D, Krueger AB (2006) Developments in the measurement of subjective well-being. Journal of Economic Perspectives 20(1):3–24
26. Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA (2006) Would you be happier if you were richer? A focusing illusion. Science 312(30):1776–1780
27. Kahneman D, Miller DT (1986) Norm theory: Comparing reality to its alternatives. Psychological Review 93(2):136–153
28. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47(2):263–292
29. Kahneman D, Wakker PP, Sarin RK (1997) Back to Bentham? Explorations of experienced utility. Quarterly Journal of Economics 112(2):375–405
30. Klein S (2006) The science of happiness: How our brains make us happy and what we can do to get happier. Da Capo Press, Tra edition, New York
31. Lama Dalai, Cutler HC (1998) The art of happiness. Riverhead Hardcover, New York
32. Layard R (2005) Happiness: Lessons from a new science. The Penguin Press, London
33. Lepper HS (1998) Use of other-reports to validate subjective well-being measures. Social Indicators Research 44:367–379
34. Loewenstein G, Schkade DA (1999) Wouldn't it be nice: Predicting future feelings. In: Kahneman D, Diener E, Schwarz N (eds) Well being: The foundation of hedonic psychology. Russell Sage, New York, 85–108
35. Loewenstein G, O'Donoghue T, Rabin M (2003a) Projection bias in predicting future utility. Quarterly Journal of Economics 118(3):1209–1248
36. Loewenstein G, Read D, Baumeister R (2003b) Decision and time. Russell Sage Foundation, New York
37. McGuire M, Raleigh M, Brammer G (1982) Sociopharmacology. Annual Review of Pharmacological Toxicology 22:643–661
38. McMahon DM (2006) Happiness: A history. Grove Press, New York
39. Medvec V, Madey S, Gilovich T (1995) When less is more: Counterfactual thinking and satisfaction among Olympic medalists. Journal of Personality and Social Psychology 69: 603–610
40. Nisbett RE, Kanouse DE (1968) Obesity, hunger, and supermarket shopping behavior. Proc Annual Convention American Psychological Association 3:683–684

41. Pavot W, Diener E (1993) The affective and cognitive cortex of self-reported measures of subjective well-being. Social Indicators Research 28(1):1–20
42. Pollak R (1970) Habit formation and dynamic demand functions. Journal of Political Economy 78:745–763
43. Putnam RD (2000) Bowling alone: The collapse and revival of American community. Simon and Schuster, New York
44. Russell B (1930) The conquest of happiness. Liveright, New York
45. Ryder HE, Heal GM (1973) Optimal growth with intertemporally dependent preferences. The Review of Economic Studies 40:1–33
46. Sahlins M (1968) Notes on the original affluent society. In: Lee R, Devore I (eds) Man the hunter. Aldine, Chicago, IL
47. Sapolsky RM, Alberts SC, Altmann J (1997) Hyper cortisolism associated with social isolation among wild baboons. Archives of General Psychiatry 54:1137–1143
48. Schkade DA, Kahneman D (1998) Does living in California make people happy? A focusing illusion in judgments of life satisfaction. Psychological Science 9(5):340–346
49. Schor J (1992) The overworked American: The unexpected decline of leisure. Basic Books, New York
50. Smith A (1759) The theory of moral sentiments. Oxford University Press, Oxford, UK
51. Smith A (1776) The wealth of nations. Reprinted by The University of Chicago Press, 1981, Chicago
52. Smith DM, Langa KM, Kabeto MV, Ubel PA (2005) Health, wealth, and happiness. Psychological Science 16(9):663–666
53. Solnick SJ, Hemenway D (1998) Is more always better? A survey on positional concerns. Journal of Economic Behavior & Organization 37:373–383
54. Stutzer A (2003) The role of income aspirations in individual happiness. Journal of Economic Behavior & Organization 54:89–109
55. Tocqueville A (1998) Democracy in America. Harper Perennial, New York
56. Tversky A, Kaneman D (1991) Loss aversion in riskless choice: A reference-dependent model. Quarterly Journal of Economics 106(4):1039–1061
57. van Praag, BMS, Frijters P (1999) The measurement of welfare and well-being: The Leyden approach. In: Kahneman D, Diener E, Schwarz N (eds) Well being: The foundation of hedonic psychology. Russell Sage, New York, 413–433
58. Veblen T (1899) The theory of the leisure class; an economic study in the evolution of institutions. Reprint by The Macmillan Company, New York
59. Wathieu L (1997) Habits and the anomalies in intertemporal choice. Management Science 43(11):1552–1563
60. Wathieu L (2004) Consumer habituation. Management Science 50(5):587–596