Aki-Hiro Sato

Applied Data-Centric Social Sciences

Concepts, Data, Computation, and Theory

Springer

Applied Data-Centric Social Sciences

Aki-Hiro Sato

Applied Data-Centric Social Sciences

Concepts, Data, Computation, and Theory

Springer

Aki-Hiro Sato
Graduate School of Informatics
Kyoto University
Kyoto, Kyoto
Japan

ISBN 978-4-431-54973-4 ISBN 978-4-431-54974-1 (eBook)
DOI 10.1007/978-4-431-54974-1
Springer Tokyo Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014942066

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

I would like to dedicate this book
to my wife, Naoko, my grandmother
and all the people who share time,
space, toils and divine favour with
me under the protection of the Trinity;
the Father, the Son and the Holy Spirit

Preface

The author has worked for 12 years as an information scientist in interdisciplinary fields involving physics, economics and information studies and has written works that specifically fall within econophysics and information sciences. The specific intention behind this book is to contribute to "econoinformatics".

The data on human societies was partial and limited at the beginning of twenty-first century. However, current data availability has improved remarkably. As a result, researchers in various fields such as economics, finance, marketing, data mining, sociology, physics and information sciences have a similar interest in data on our society and study societal issues using a large amount of these data. They have opened up a new paradigm of studies on society that is described by a single keyword: data. In this book, these new emerging fields are termed "Applied Data-Centric Social Sciences", which in association with data use, and this book has been written to share a vision of these fields.

Human society often shows interesting properties, such as non-stationarity, synchrony and spatiotemporal patterns. In order to capture these properties, individual behaviour, social relationships among individuals and man-made artificial systems need to be understood.

Since the nineteenth century, social scientists have conducted empirical investigations to understand the characteristics of collective versus individual behaviour and relationships between individuals within a society. Concurrently, data on socio-economic technological systems have accumulated in various fields of study.

Since research topics within applied data-centric social sciences are wide and deep, this book attempts to introduce some fundamental segments of these fields, including several mathematical expressions and some techniques to handle a vast amount of data and computer analysis. This work is also based on several example studies of data-oriented investigation in which advanced mathematics is used to analyse and model several specific problems.

The fundamental philosophy underlying this book is that both mathematical and physical expressions should be used to express actual, real-world data with high accuracy and thereby understand data-generating mechanism.

In data-centric science method, thinking first starts from the data in a specific field. Next, a search is attempted for an adequate method or expression to investigate the data. Explanatory data analysis provides an improvement cycle

through data acquisition, data collection and data analysis to reach interpretation. This type of activity constructs a PDCA (plan-do-check-action) cycle, which is a part of data-centric science. Data preservation and data recycling may thus be examined over a longer time horizon, a process sometimes referred to as 'data curation'.

Mathematics are useful tools to express processes and states of our socio-economic technological systems. If problems can be described by well-defined mathematical notations and modelled by expressions, then cyber-enabled techniques can be constructed—such as automated data collection, automatic data verification and optimisation techniques—by using both mathematical and physical expressions and models.

Data is defined as several numbers to describe physical quantities (e.g. length, weight, time and velocity) or a number of texts expressing the actual situations or processes. The data is collected to trace real-world situations (records) to transfer information on what is focused on in the real world to other people (communications). Why should the data be analysed? The dominant purpose of data analysis is deeply associated with decision-making. In general, humans want to know and understand processes and phenomena in more details when they have to make some decisions in actual environments where they live. Thus, the results of data analysis are used as information for decision-making in the real world.

Some contact points to actual society are present in data analysis on socio-economic technological systems. Legal issues on data-centric social sciences are also addressed, although some of these issues are still under discussion.

Furthermore, this book contains separate intentions behind each chapter. The first of these is how to describe mathematical and statistical methods for data analysis. The second is to look at the background of data-generating mechanism. The third is the motivation for applying data-centric sciences to socio-economic technological systems. The example studies using the data of a specific field are presented to help readers to understand the different situations of socio-economic technological systems and how mathematical and physical methods are applied to actual data. Activities in data-centric social sciences are addressed as much as possible.

To satisfy these intentions, this book is organised as follows. Chapter 1 discusses the concepts used to deal with data on socio-economic technological systems and the reliance of the data-centric social sciences on the data for social activities, relationships and behaviour. Chapter 2 explains the research framework of applied data-centric social sciences. Why do scientists acquire, collect and analyse data on human society? How can the results of data analysis be utilised? This chapter also discusses methodologies within applied data-centric social sciences.

Chapter 3 introduces mathematical expressions used to describe societies, human behaviour and relationship. Several fundamental methods are explained, including statistical procedures, stochastic methods, network description and geographic information. Chapter 4 shows several methods of processing data with

computers. Database servers and parallel computation techniques are needed to handle large scale data.

Chapter 5 shows an example study of risk assessment in the foreign exchange market by using both q-Gaussian and Pearson type IV distributions, while Chap. 6 discusses a method to quantify states of the foreign exchange market by using a recursive segmentation procedure. Chapter 7 presents analysis of Japanese hotel booking data and quantifies the regional dependence of hotels, and Chap. 8 looks at relationships between flight ticket prices and their geodesic distance. Chapter 9 considers the relationship between electric power consumption per capita and economic performance (GDP per capita). Finally, Chap. 10 examines the future of the applied data-centric social sciences.

This book hopes, in this way, to encourage readers to acquire, collect, store, analyse and interpret data from socio-economic technological systems, in order to solve their own problems.

Kyoto, Zurich, 2013, 2014 Aki-Hiro Sato

Acknowledgments

I would like to express my sincere gratitude to Dr. Hideki Takayasu (Sony Computer Science Laboratories, Tokyo, Japan), Professor Misako Takayasu (Tokyo Institute of Technology, Tokyo, Japan) and Dr. Yasuji Sawada for their kind support during my opening as a researcher. It was pleasure to learn fundamental knowledge and professional methodologies on physics and information sciences at Tohoku University, Sendai, Japan.

I am thankful to Professor Thomas Lux (University of Kiel and Institute for World Economy, Kiel, Germany) Professor Burda Zdzisław (Jagiellonian University, Krakow, Poland), Professor Janusz Hołyst (Warsaw University of Technology, Warsaw, Poland), Professor Dirk Helbing (ETH Zurich, Switzerland) and Professor Frank Schweitzer (ETH Zurich, Switzerland) for our stimulating discussions.

Nothing gives me greater pleasure than to meet and discuss many things with Professor Janusz Hołyst. His interests cover applications of physics to socio-economic phenomena. My first visit to overseas University was Warsaw University of Technology in Poland. Professor Janusz Hołyst kindly invited me to his University, and I had an opportunity to stay at Warsaw from April to May 2007.

The next chance to find a new method for economics was obtained from German economics. I had the opportunity to stay at the University of Kiel from March to September 2010 (JSPS Excellent Young Researcher's Oversee Visiting Programme). Professor Thomas Lux kindly prepared my visit to his University with my wife and children. I learnt a lot from him and his colleagues. Discussions with Dr. Jaba Ghonghadze and Dr. Tae-Seok Jang were fruitful. It was always exciting for me to find methods for economics and statistics at the University of Kiel.

Professor Dirk Helbing kindly invited me to VISIONEER Workshop held in ETH Zurich in January 2010. Thanks to Professor Frank Schweitzer, I had the opportunity to stay at ETH Zurich in October 2012. I had one more chance to participate in the Swiss-Kyoto Symposium held at ETH Zurich in November 2013. I obtained a great deal of insight from Professor Dirk Helbing and Professor Frank Schweitzer.

I stayed at Jagiellonian University in March 2012 and discussed random matrix theory and its application to econophysics with Professor Burda.

I learnt much from them all, and they gave me an understanding of the importance of data-driven investigation of socio-economic technological systems.

It has also been invaluable to discuss finance, economics and econophysics with Professor Yuji Aruka (Chuo University, Tokyo, Japan), Professor Mieko Yamawaki-Tanaka (Tottori University, Tottori, Japan) and Professor Hideaki Aoyama (Kyoto University, Kyoto, Japan) in workshops at Yukawa Institute for Theoretical Physics and Institute of Statistical Mathematics. Further, it was my pleasure to collaborate with Professor Takaki Hayashi (Keio University, Yokohama, Japan). It was beneficial to discuss agent-based modelling with Professor Akira Namatame (National Defence Academy of Japan, Yokosuka, Japan) and Professor Shu-Heng Chen (National Chengchi University, Taipei, Taiwan). Furthermore, I would like to express my sincere gratitude to Dr. Hidefumi Sawai (National Institute of Information and Communications Technology) for his guidance.

I would also like to acknowledge my debt for fruitful discussions with Mr. Shigehiro Kato (Recruit) in Chap. 1; Mr. Takashi Isogai (Bank of Japan) in Chap. 5; with Mr. Masaki Nishimura (National Statistics Centre), Mr. Toshimi Yamada (National Statistics Centre), Mr. Kohichiro Furuichi (Statistics Bureau, Ministry of Internal Affairs and Communications in Japan) and Mr. Kazumasa Matsushita (Statistics Bureau, Ministry of Internal Affairs and Communications in Japan) in Chaps. 3 and 4; with Mr. Kotaro Sasaki (Jalan), Mr. Daichi Tanaka (Jalan) and Mr. Hiroshi Yoshimura (Jalan) in Chap. 7; and with Ms. Youko Miura (AB-ROAD) in Chap. 8. I wish to further express my sincere gratitude to Mr. Stefano Balietti and Dr. Olivia Woolley for helping me to prepare the manuscript for Chap. 3, to Mr. Yuzo Morita and Mr. Hiroshi Kajikawa for their assistance in writing drafts of Chaps. 3 and 5. I am further thankful to Mr. Minoru Noda for helping me to prepare the manuscript for Chap. 7. Mr. Yutaka Hirachi assisted me in preparing this book.

The contents of this book are based on research financially supported by the Grants-in-Aid for Young Scientists (B) (#23760074), by the Japanese Society of Promotion of Science (JSPS) and the Grants-in-Aid for Scientific Research (C) (#25390152).

Contents

Acronyms

3D	Three-Dimensional
AIC	Akaike Information Criterion
AMR	Automated Metre Reading
API	Application Programming Interface
AR	AutoRegressive (model)
ARCH	AutoRegressive Conditional Heteroscedastic (model)
ARMA	AutoRegressive Moving Average (model)
BI	Business Intelligence
BIC	Bayesian Information Criterion
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
CEP	Complex Event Processing
CG	Conjugate Gradient (method)
CPS	Cyber-Physical System
CSD	UN Commission on Sustainable Development
CSDIs	UN Commission on Sustainable Development Indicators
CSV	Comma-Separated Values
DDEMS	Decentralized Distributed Energy Management System
DMA	Direct Market Access
DQM	Data Quality Management
EC	Electronic Commerce
e-commerce	Electronic commerce
EHR	Electronic Health Record
EIA	U.S. Energy Information Administration
ES	Expected Shortfall
eWoM	Electronic Word of Mouth
GARCH	Generalised AutoRegressive Conditional Heteroscedastic (model)
GBMM	Generalised Box–Müller Method
GDP	Gross Domestic Product
GIS	Geographical Information System
GNP	Gross National Product
HPC	High Performance Computing
IATA	International Air Transport Association
IBRD	International Bank for Reconstruction and Development

ICAO	International Civil Aviation Organisation
ICT	Information and Communication Technology
IE	Information Extraction
IEA	International Energy Agency
IoT	Internet of Things
IR	Information Retrieval
ISO	International Standardisation Organisation
ITS	Integrated Transport Systems
ITU	International Telecommunications Union
KF	Kalman Filter
KISS	Keep It Simple and Straightforward (principle)
Libor	London Interbank Offered Rate
M2M	Machine-to-Machine
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
OI	Optimal Interpolation
OLS	Ordinary Least Squared (regression)
PDCA	Plan-Do-Check-Action
PDF	Probability Density Function
R&D	Research and Development
RAID	Redundant Arrays of Independent Disks
RBC	Real Business Cycle
RDBMS	Relational Database Management System
RMA	Reduced Major Axis (regression)
SARS	Severe Acute Respiratory Syndrome
SDIs	Sustainable Development Indicators
SGD	Stochastic Gradient Descent
SQL	Structured Query Language
TSV	Tab-Separated Values
UN	United Nations
UNSD	United Nations Statistical Databases
UTC	Coordinated Universal Time
VaR	Value at Risk
VGI	Volunteered Geographic Information
WCC	Weakly Connected Component
WGS 84	World Geodetic System 1984
XML	Extensible Markup Language

Notations

m	The number of parameters
T	The number of observations
$l(\theta_1, \ldots, \theta_m)$	The loglikelihood function (or $L(\theta_1, \ldots, \theta_m)$)
θ_i	The i-th parameter
$\hat{\theta}_i$	The i-th parameter estimate
A_{ij}	An element of the i-th row and the j-th column of an adjacency matrix in a network
a	The power law exponent
f	A function
A	A set
a_1	An element of a set A
a_2	An element of a set A
$I(p_1, p_2)$	Kullback–Leibler divergence
n	The number of particles in Ising model and the number of people in Granovetter threshold model
E	A set of nodes in a graph
e_i	A node
V	A set of links in a graph
l_i	A link
N	The number of nodes in a network
L	The number of links in a network
$\langle l \rangle$	The mean path length
$\langle k \rangle$	The average degree of the neighbourhood of each node
d	A density of a network
k_i	A degree of the i-th node
$k_i^{(in)}$	An in-degree of the i-th node
$k_i^{(out)}$	An out-degree of the i-th node
$c(e_i)$	A node degree density of the i-th node
$c^{(in)}(e_i)$	An in-degree node density of the i-th node
$c^{(out)}(e_i)$	An out-degree node density of the i-th node
$C_D(e_i)$	A centrality of the i-th node

$C_D^{(in)}(e_i)$	An in-degree centrality of the i-th node
$C_D^{(out)}(e_i)$	An out-degree centrality of the i-th node
$H(\boldsymbol{A})$	Network entropy of a network described by an adjacency matrix $\boldsymbol{A}$
r	Assortativity
q_{ij}	The fraction of links in a network that connect a node of type i to one of type j
Q	Modulatily
K	The number of community
ϕ°	Latitude of geographic coordinate system measured in degrees: $-90^\circ \leq \phi^\circ \leq 90^\circ$
λ°	Longitude of geographic coordinate system measured in degrees: $-180^\circ \leq \lambda^\circ \leq 180^\circ$
ϕ	Latitude of geographic coordinate system measured in radians: $-\pi/2 \leq \phi \leq \pi/2$
λ	Longitude of geographic coordinate system measured in radians: $-\pi \leq \lambda \leq \pi$
D	Distance
J	Interaction constant of the Ising model
θ	A social temperature of the Ising model
σ_i	A state variable of the Ising model
$p(\sigma, t)$	The individual probabilities for the state σ
$p(x)$	The probability density function (PDF) of x
$\ln x$	The logarithm to the base e in terms of x
$\log_{10} x$	The common logarithm in terms of x
e	Euler's number (approximately 2.718281828)
w_{ij}	The weight matrix of spatial autocorrelation
I	Spatial autocorrelation the Moran's I
C	Spatial autocorrelation the Geary's C
G	Spatial autocorrelation the Getis's G
$\mathrm{Var}[X]$	The variance of x
$\mathrm{E}[X]$	The expectation value of x
λ_3	The skewness
λ_4	The kurtosis
σ	The standard deviation
m_1	The mean
m_m	The median
m^*	The mode
$\mathrm{Cov}[X, Y]$	The covariance between X and Y
$\mathrm{Corr}[X, Y]$	The correlation between X and Y
$\gamma_{s,t}$	The autocovariance function
$\rho_{s,t}$	The autocorrelation function
VaR_c	$100(1-c)$ % value at risk

ES_c	$100(1-c)$ % expected shortfall
$\boldsymbol{A}^{\mathrm{T}}$	A transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	An inverse matrix of a matrix $\boldsymbol{A}$
$\delta(x)$	A Dirac's delta function
δ_{ij}	A Kronecker's delta function

Part I
Introduction

Chapter 1
Introduction

Abstract Recent development of information and communication technology enables us to acquire, collect, analyse data in various fields of socioeconomic-technological systems. In this chapter, we will address data from several different perspectives and define the applied data-centric social sciences. I will explain that limitation of our ability to understand our society from inductive approach is origins of complexity. Concepts and methodologies of data-centric science will be introduced and their potential applications and existing studies will be mentioned.

1.1 Why Write this Book?

The purpose of this book is to share my vision of applied data-centric social sciences. I suggest that applied data-centric social sciences can be defined as a transdisciplinary study of our society based on data from socioeconomic-technological systems, using methodologies from data-centric science. Data-centric science is an emerging field related to "data" in various fields of sciences. Our society is located in nature and can be understood as a human-machine system. Therefore, studies on our society should include topics of both human nature and man-made objects, such as technology.

Traditionally, social sciences are defined as the academic disciplines concerned with society, individual behaviour and the relationships among individuals within a society. This is a common umbrella term covering anthropology, economics, political science, psychology and sociology. Strictly speaking, the social sciences are also divided into two types of disciplines:

- empirical social science
- normative social science

Empirical social science attempts to investigate human behaviour through observations or experiences. Economics, sociology, history and geography mainly focus on empirical research. While normative social science, such as nomology, logic, laws, ethics, and aesthetics, attempts to clarify correct meanings and the validity of

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_1, © Springer Japan 2014

Table 1.1 An example of classification of social sciences

	Norms	Politics/power	Market
Mechanism	Law	Politics	Economics
Phenomena/analysis	Sociology	Policy making	Marketing
Application	Social welfare	Education	Management

social contexts. In fact, there is a clear difference between empirical and normative social sciences. For example, laws of empirical social sciences are common phenomena observed in our society with a high probability of validity, while laws of normative social sciences are rules that we should follow in our society. The model of the empirical social sciences is a simple description of our society or individuals, which is constructed in order to understand phenomena observed in a socioeconomic-technological system. In contrast, the model of the normative social sciences is an ideal realisation of our society. If we behave according to this model, we expect that our organisation or our behaviour can be formed following the model assumptions. We can also classify several disciplines that belong to social sciences as shown in Table 1.1. In this table, there are two aspects, consisting of fields (norms, politics/power, and market) and methodologies (mechanism, phenomena/analysis, and application). Our concept of applied data-centric social sciences influences ways to study in each field. We can, of course, understand the social sciences from other aspects.

The original empirical science of society was established in 19th century by several researchers. Max Weber is one of the most famous founders of empirical social science. He proposes methods of empirical social science and classified several types of laws observed in social phenomena.

In his approach, he mentions two kinds of meanings: the subjective meaning of the actual existing agents and the theoretically assumed pure type of subjective meaning by hypothetical agents. He also regards two kinds of understandings as important: the direct observational understanding of the subjective meaning of a given act (both verbal and nonverbal actions) and direct observational understanding of irrational emotional reactions. Empirical science attempts to interpret meaning with clarity and to ensure the verifiable accuracy of an insight and understanding. The line between meaningful action and merely reactive behaviour without subjective meaning, however, cannot be clearly drawn in an empirical sense.

To understand society from an empirical point of view, we can use three approaches:

1. an historical approach to the meaning of concrete individual action
2. an observational approach to socioeconomic mass phenomena
3. a model approach to scientifically-formulated pure type of socioeconomic-technological phenomena observed, with high probability of validity.

Recent advanced computerising society has enabled data-driven research on our society, human behaviour and the relationships among individuals. Before the computerisation of our society, observations of and experiments on human nature were

strictly partial. Only small-scale experiments in a laboratory and direct/indirect questionnaire surveys were possible. Frequency and coverage of data on socioeconomic-technological systems were also partial. Traditionally social and cultural data are collected via field studies, user panels, focus groups, interviews, questionnaires and surveys (small or huge like the decennial U.S. census).

However, recent computer networks consisting of computers and both wired and wireless networks allow the automatic detection of human behaviour and accumulation of data on socioeconomic-technological systems.

This appears to have created a kind of transition phase from quantity to quality. Rich data is slowly changing the methodology used in social sciences. Since data-centric social sciences are classified as a part of data-centric science, they belong within both inductive and cyber-enabled social sciences (see Sect. 1.4). The three types of empirical approaches mentioned above can be improved both quantitatively and qualitatively. Of course, the importance of cyber-enabled methodologies in social sciences has increased as the data has expanded.

There are, of course, two perspectives on data-centric social sciences: empirical and normative. The research environment from the empirical perspectives has been drastically improved by the computerisation of our society. Increasingly, the social and cultural interactions passing through or taking place on the web are considered as valuable sources of data for social and cultural research [127, 170]. New methodologies of social sciences based on data have emerged.

Specifically, query log analysis of a Web search engine reveals human typical behaviour [14, 100]. For example, one can watch and predict the spread of flu [69, 77], unemployment rates [10, 46, 113], trading volume and return volatility of financial markets [25, 43, 164, 185, 208], travel tendencies [155, 220], private consumption tendency [209] and results of elections [191]—all from search queries. However, the accuracy of these predictions is still under discussion [29].

Meanwhile, computerisation of our society creates issues that we have not considered from a normative point of view. How do we behave in cyberspace? How do we establish our organisation within a network structure? How do we publicise information in both the real world and cyberspace? We can easily address several normative issues; however, several important issues are still under discussion. Legal issues of data-centric researches (secondary usage of data and data protection acts) are crucial today. The trust formation of human relationships [27], analysis of complex social phenomena [87] and design of societal framework [103] have been studied from a normative point of view.

Thus, fields of data-centric social sciences are slowly being established following the spread of computers and the Internet. In the 1960s, computers began to play an important role in business. Companies started to introduce mainframe computers in their offices. At the end of the 1970s, personal computers changed work and business. In the 1980s, companies began to handle consumer information easily, due to the development of relational databases. Data analysis in business was often used.

The origins of the Internet can be traced back to the development of a packet switched network called ARPANET by the U.S. Federal Government's Defence Advanced Research Projects Agency (DARPA). In 1969, this agency started

building ARPANET supporting various computer science and military research projects. By 1980, a stable protocol suite was established, facilitating connectivity among ARPANET computers. In 1983, transmission control protocol/Internet protocol (TCP/IP) started to be used by all ARPANET users. The Internet has been used commercially since 1992, after the removal of a ban on the use of the Internet in commercial sectors. Many applications and services have been designed and realised based on combinations of several types of Information and Communication Technology (ICT). The Internet consists of network architectures (both wired and wireless communication), computer hardware, and cyber-enabled services and applications [48, 96].

Recent ICT has advanced so far that it is indistinguishable from *magic*. This is known as the Clarke's Third Law. Arthur C. Clarke mentioned this law in his famous essay entitled "Hazards of Prophecy: The Failure of Imagination" in Profiles of the Future (1962) [41]. Since advanced technologies are constructed of many elements, which are complex systems themselves, ordinary people cannot completely identify these mechanism, even professionals. Therefore, it is often perceived as a kind of magic.

Nonetheless, there is no magic in the world. We just perceive something as magic, even though the magic is always produced by a combination of operational methods. However, this magic actually plays a significant role and influences many of those who perceive it as a magic. Therefore, I believe that we need to possess adequate knowledge, morals, and techniques in order to establish the stable world based on advanced technologies (magic). Specifically, to know the applicable limits of technology is important, as well as to understand how to apply the technology to actual problems.

Recently, people can access data in various fields due to developments in ICT. This circumstance enables us to study our society based on a large amount of data on socioeconomic-technological systems. One can collect and accumulate large amounts of socioeconomic-technological data on human activities, and then analyse and visualise them. Vast amounts of data collected from socioeconomic-technological systems have allowed new types of commercial services and research fields to emerge.

Data-centric social sciences have been recently developing based on ICT. A large amount of data on socioeconomic-technological systems has slowly accumulated in several institutions and fields. The data is generated and collected in some type of data-generating mechanism and then stored in a database or computer storage. The data is eventually distributed or analysed for a purpose, which is to interpret the world from the data and to make decisions. We can also start to formulate a model of a specific phenomenon from the data since the data may provide us with useful insights to construct a model. For this reason, this book is needed to share this vision of applied data-centric social sciences.

There is a pipeline from data to our decision-making [37]. The study of decision making has a long, distinguished, and interdisciplinary history. According to Knight [119], we can distinguish between risk and uncertainty. Risk defines decision situations in which the probabilities are objective or given, such as betting on a flip

of an evenly-weighted coin. Uncertainty defines situations in which the probabilities are subjective (the decision maker must estimate or infer the probabilities). Examples of this type of situation are the decision to invest overseas, launch a new product, and to buy or sell stocks. Data analysis can contribute to estimating the probabilities of uncertainty.

This is not a single process but construct a cycle consisting of data and thinking. In methods of data-centric science, we firstly start our thinking from data in a specific field. Next, we attempt to search an adequate method or expression to investigate the data. Explanatory data analysis provides an improvement cycle from data acquisition, data collection, data analysis and interpretation. Such a type of activity constructs a PDCA (Plan-Do-Check-Action) cycle. I therefore start this chapter with an overview of our society.

1.2 Overview

1.2.1 Our World

The problem of socioeconomic-technological systems is deeply related to human. Since the populations determine both economic and social affairs mainly, it seems to be worth grasping populations and their spatial distribution.

What do the data on populations tell us? First of all, let us address the largest cities in the world. Figure 1.1 shows the geographical position of the 135,074 largest cities around the world. The data source is geonames (http://www.geonames.org), which is an open database of population in cities with geographical information. Figure 1.2a shows the ranking by population of the largest 50 cities. The largest city of the world is Shanghai, China, with a population estimated at 14,608,512, according to the figures for March 2012. The second largest population is in Buenos Aires, Argentina. The third is Mumbai, India. The largest urban areas in the world is Tokyo-Yokohama, Japan, with a population estimated at 37,239,000 in 2013 [49]. The world population reached seven billion on October 31, 2011 according to an estimate by the United Nations Population Found [18]. According to the data on world population issued by the Population Division in the Department of Economic and Social Affairs of the United Nations, the world's population is predicted to reach eleven billion in 2100, as shown in Fig 1.2b.

How does each person behave in each country? In fact, there are several dimensions used to quantitatively measure human behaviour. The characteristics of averaged behaviour can be derived from World Development Indicators provided by the World Bank's World DataBank.[1] One of the best approaches to understanding of our society is to consider allometric relationship in analogy of biology.

[1] The World Bank's World DataBank: http://data.worldbank.org.

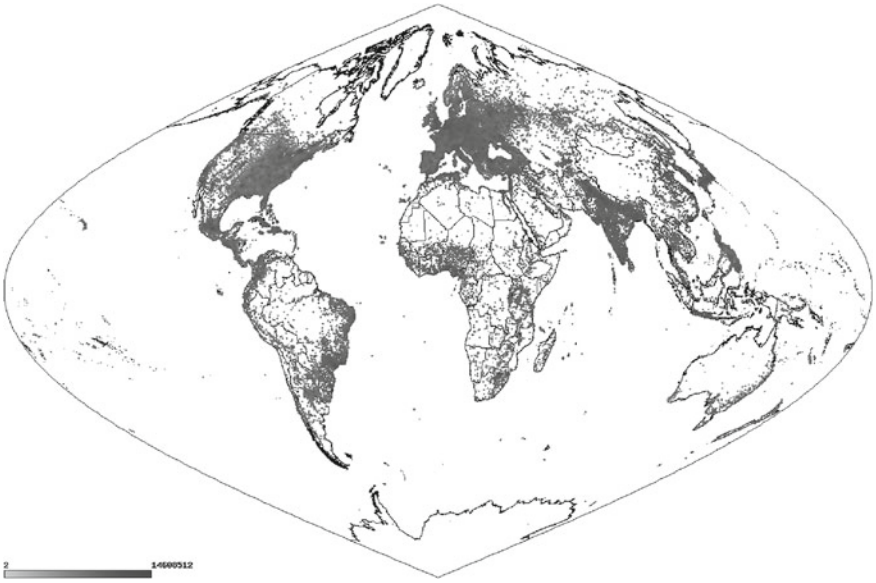


Fig. 1.1 Geographical positions of around 135,074 of the largest cities around the world. This image is drawn by using data on major city population recorded in 2012. The *dark grey* represents cities with large populations. The grey represents those with smaller populations

Allometry is a relationship seen between two physical quantities of living things. For example, it is well-known that there is a relationship between body size of creatures and their mass [95, 106, 210]. An elephant is both heavy and big, but a mouse is both light and small. If we plot the relationship between weight of each creature and its representative length, then they have a positive correlation.

Allometric relationship can be also seen in our society [19, 214, 225]. There are scaling relationships between urban indicators and city size [19, 225]. Several urban indicators such as GDP, total electrical consumption, total housing, total employment, road surface and so forth, as the populations increase. The data on several types of human behaviour also show some correlations, including the relationship of some physical quantities each person consumes or produces to his or her economic contribution as an individual (Gross Domestic Product (GDP) per capita).

Figure 1.3 shows the relationship between GDP per capita (USD in 2013 per capita) and energy consumption per capita (kg of oil equivalent per capita) in 2000 and 2010. Each point represents the relationship in each country. Developing countries are positioned on the left hand side and developed countries are found on the right hand side, which implies that energy usage per person is strongly correlated with economic activity per capita. This implies that the quality of our life is generated by the contributions of man-made systems. Figure 1.4 shows the relationship between GDP per capita (USD in 2013 per capita) and CO_2 emissions per capita (metric tons per capita) in 2000 and 2010. We can see a clear power law relationship between

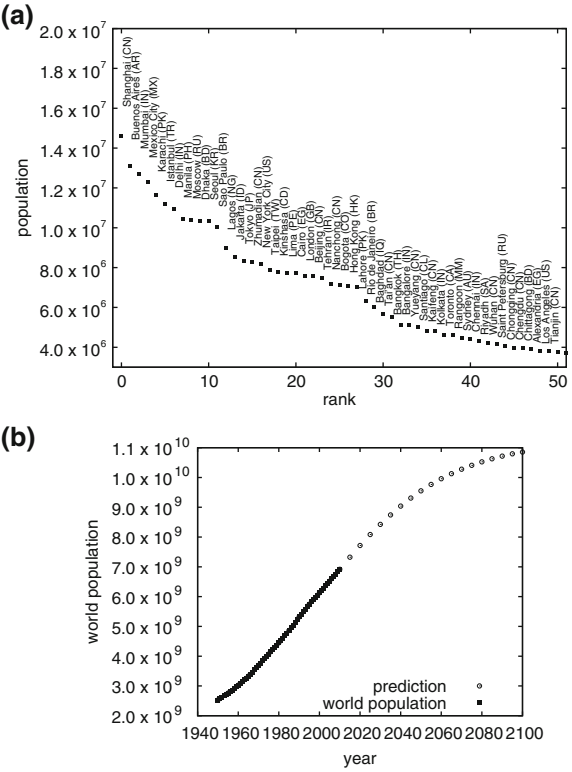


Fig. 1.2 **a** The ranking by population of the largest 50 cities as of 2012. The data source is geonames. **b** The total population (both sexes combined as of 1 July) issued by the population division in the department of economic and social affairs of the United Nations in 2012

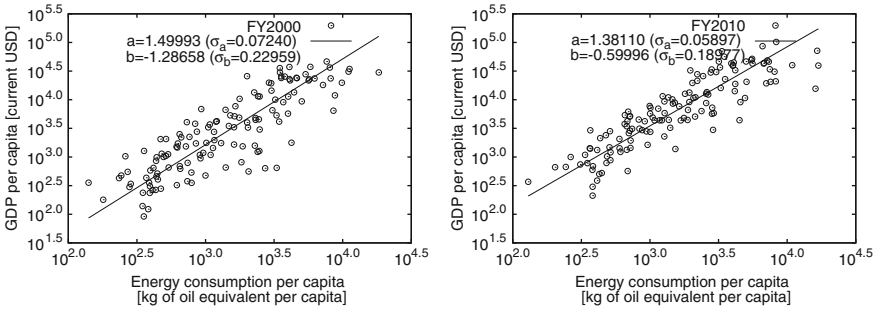


Fig. 1.3 The relationship between gross domestic product (GDP) per capita and energy consumption per capita. World development indicators by World DataBank of the World Bank are used. Unfilled circles represent the relationship of each country in (*left*) 2000 (133 countries and territories) and (*right*) 2010 (133 countries and territories). A *solid line* presents the power law relationship between GDP per capita and energy consumption per capita

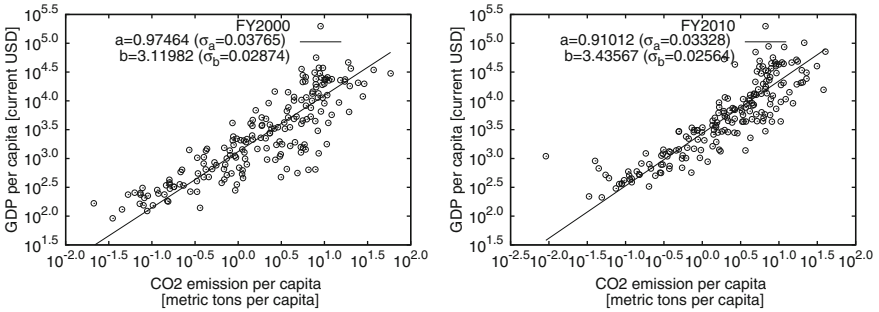


Fig. 1.4 The relationship between GDP per capita and CO_2 emissions per capita. World development indicators by World DataBank of the World Bank are used. Unfilled circles represent the relation of each country in (*left*) 2000 (188 countries and territories) and (*right*) 2010 (185 countries and territories). The *solid line* presents the power law relationship between GDP per capita and energy consumption per capita

GDP per capita and CO_2 emissions per capita. Thus, CO_2 emissions also show an obvious correlation with economic activity.

Studies in allometry often assume that the relationship between the two measurable quantities y and x is often expressed as a power law [225]:

$$y = cx^a, \tag{1.1}$$

where a is the scaling exponent of the law, and c a positive constant. The parameters are estimated by using the reduced major axis (RMA) regression (see Sect. 3.1.9.2) for its logarithmic form:

$$\log_{10} y = a \log_{10} x + \log_{10} c. \tag{1.2}$$

Figure 1.5 shows the evolution of the power law exponent for the relationship between GDP per capita and energy consumption per capita and for the relationship between GDP per capita and CO_2 emission per capita by year. It is confirmed that the power law exponent for the relationship between GDP per capita and energy consumption per capita showed smaller values than before and after the period of 2004–2008. Meanwhile, the power law exponent for the relationship between GDP per capita and CO_2 emission per capita decreased. This seems to correspond to a difference between developed countries and developing countries in how mind-set changed individual behaviour during the globalisation of the world economy.

Other types of allometric relationships have been observed in the context of societal issues. For example, economic indices can be seen as a function of urban population, and level of energy consumption are also associated with urban population. The urban infrastructure or functions, including roads and railroad, is associated with national populations [19, 225].

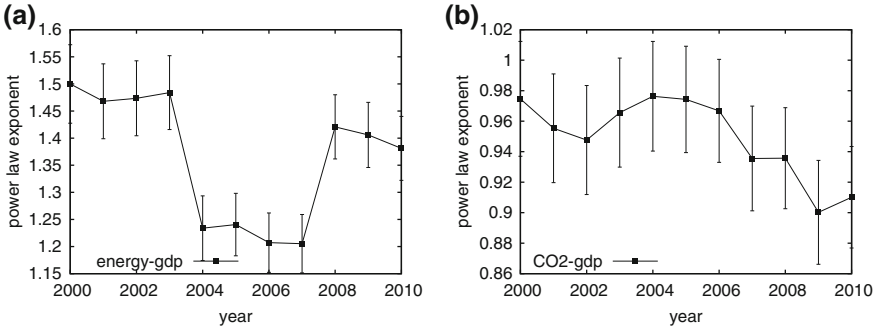


Fig. 1.5 The power law exponent for the relationship **a** between GDP per capita and energy consumption per capita and **b** between GDP per capita and CO_2 emissions per capita for the period of 2000–2010. The error bars represent standard errors of a parameter estimate

1.2.2 Origin of Complexity

We can extract information from a collection of data, construct knowledge from plentiful information, and hopefully extract wisdom from several pieces of knowledge. Specifically, researchers in the fields of sociology, economics, informatics, and physics are focusing on these frontiers and have launched data-driven investigation for our society in order to understand the complexity of socioeconomic-technological systems. This is an inductive approach from facts to wisdom.

However, since our society which is the sum total of both internal and external states of individuals, is several orders of magnitude more complicated than each individual, it appears to be difficult to image how we can manage to capture the real totality of the state of society from the cooperation of many agencies. The nature of this problem is referred to as "complexity", which is a new research field into understanding how groups of people, organisations, communities, and the economy actually behave in the real world.

According to von Foerster [211], complexity is not a property that observed systems possess; rather, it is to be perceived by observing systems. He asks us about it through the following question: *Are the states of order and disorder states of affairs that have been discovered, or are these states of affairs that are invented?* If states of order and disorder are discovered, then complexity is a property of the observed systems. If invented, then it pertains to the observing systems. Foerster's definition of complexity proposes that the relative degree of order and disorder is determined by the degrees of freedom within an observed system and an observing system. One of the most significant reasons why we recognise complexity in observed systems is because the ability with which we are able to observe the systems is finite, and our memory and *a priori* knowledge are limited. These limits lead us to our bounded rationality, or ignorance.

1.2.3 Big Data and Landauer's Experiment

How many megabytes of memory does our brain have? Von Neumann proposes that human memory can be estimated as 100 Eb from all the neural impulses conducted in the brain during a lifetime [212]. Another method is to estimate the total number of synapses, and then presume that each synapse can hold a few bits. Estimates of the number of synapses have been made in the range from 10^{13} to 10^{15}, with corresponding estimates of memory capacity. This estimation suggests that human memory capacity is from 10 Tb to 1 Pb.

Furthermore, there is another approach to estimating the human memory capacity. How fast can we interpret information from physical stimuli? Landauer studied how much people remember at Bell Communications Research [126]. The remarkable finding of this study was that human beings remember very nearly two bits per second under all experimental conditions (visual, verbal, and musical).

How many symbols can we remember in our life? The 35-year accumulation of a human beings' memory is estimated at 0.2–1.4 Gb if a loss of memory is assumed. If the amount of data is greater than this estimate of memory capacity, then we cannot memorise all the data even over an entire life.

Clearly, almost all the data that we deal with in our current, advanced-information society is far beyond our cognitive capacity. All we can do is to construct a system to handle the large amount of data exceeding our human ability. Recently, this type of data is often called "big data". The definition of big data is still under discussion, but the most famous definition is constructed from the three perspectives of data volume, data velocity and data variety (3V).

From the data-centric viewpoint, complexity is also defined from these three perspectives; volume, velocity, and variety. Volume means that we recognise complexity in the data when the amount of data is larger than a single human's cognitive capacity (0.2–1.4 Gb). The second perspective, velocity, considers the data-generating speed. We recognise that the data is complex when the data generating speed exceeds our cognitive speed (2 bps). The third perspective focuses on the number of element types referred to in the data. We realise that the data is complex when the number of types is larger than the number of symbols which we know (depending on their own experiences). In these three dimensions, complexity of data seems to be related to our cognitive capacity.

We can measure the degree of big data by using 3V. Figure 1.6 shows a scale for data on the three dimensions of 3V. Each dimension has a certain scale. The volume is measured by bits or bytes (=8 bits). Since the memory of human and a learning speed is finite, in fact it is possible for us to remember 100 bytes of text data, but it is hard to memorise a gigabyte of text data. It is more difficult to handle a petabyte of text data when using even an up-to-date personal computer. The velocity means an average period of data generation. Daily data can be easily retained; however, it is hard to recognise data by the second. Variety expresses the number of items or elements comprising the data. Thus, the area of triangle constructed from these three

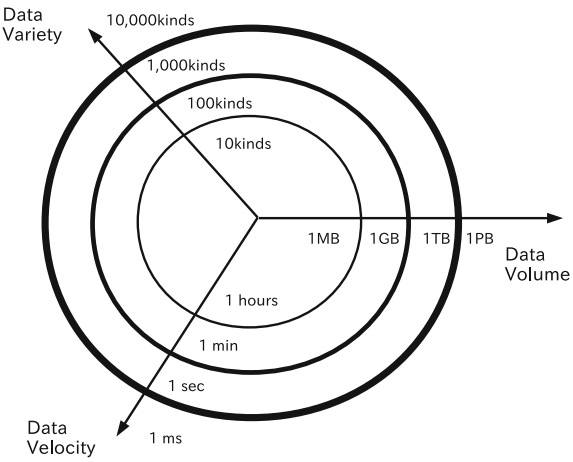


Fig. 1.6 Pictorial illustration of the three dimensions of data. Big data expands on all three fronts of each perspective

dimensions in Fig. 1.6 corresponds to the capacity that is needed when we analyse the data with the defined properties.

In this sense, we also need to carefully consider methods of acquisition, analysis, curation, storage and usage of the vast amount of data in our computer system. Big data analytics is data-centric research based on this vast amount of data. Generally speaking, the gigabytes level of text data cannot be analysed by using personal computers. Presumably we need to use parallel computation techniques or distributed architecture in order to analyse this level of data. Thus, it is possible to deal with big data with cyber-enabled methodologies: digital data and computation.

In addition to 3V, human nature makes our society more complicated. Observations sometimes influence the behaviour of observed individuals. Furthermore, the principle of the uniformity of society seems to be weak, since human responses to stimuli or information are not unique.

In this book, I present example studies of observation and data analysis with a large amount of data in several socioeconomic-technological systems. I suggest that some of them should be placed within "econoinformatics" in applied data-centic social sciences. The applied data-centic social sciences not only need powerful computer systems but also mathematical methods, including analytics, geometry, statistics, time series analysis, network analysis and so forth. This will provide us with useful methods to visualise and quantify the behaviour of human beings, and will provide deep insights for us to adapt our environment as it evolves over time.

In addition, we should mention the problem of data linkage in data-centric studies. Data linkage is defined as a method of linking data from different sources with the same elements. Integrated data from different data sources can provide us with new insights more than each data source. Comprehensive study is further useful in obtaining new findings on our environment. This technique is sometimes mentioned in studies of data integration and data fusion [91, 112, 161].

1.3 Data

1.3.1 What is Data?

Data is a collection of facts that can be described as values or expressions. Data is expressed as a set of descriptions or numbers and can be qualitative or quantitative. Qualitative data is descriptive information, whereas quantitative data is expressed by numerical information (numbers). In this book, I mainly focus on quantitative data since cyber-enabled technique is adaptable. Methods of qualitative data analysis are introduced in Richards's book *Handling Qualitative Data* [168].

Figure 1.7 shows the fundamental research framework of data-centric science. Data is collected from the real world. Recently our society can often be computerised, and many computers are used to facilitate social services. These computers are connected with one another through a high speed network infrastructure. The data used in data-centric social sciences are generated from personal computers, smart phones, telephones, and sensors. Electronic commerce (EC) systems can also be used as a data source. Since goods or services in EC are provided through Internet applications or Application Programming Interface (API), we can accumulate data on our society. Several public institutions provide open data for social sciences. For examples, e-Stat is a portal site of the Official Statistics of Japan, developed by Statistics Bureau of the Ministry of International Affairs and Communications.[2] UK data service also provides the data that promotes evidence-based social research and policies.[3] The European Commission provides a portal site of European statistics.[4] Data.gov provides public access to machine readable datasets generated by the Executive Branch of the Federal Government of the United States of America.[5] In the web-page of open data index[6] of the Open Knowledge Foundation,[7] we can find ten types of open data:

- Transport timetables
- Government budget
- Government spending
- Election results
- Company register
- National map
- Legislation
- Postcodes/Zip codes
- Emissions of pollutants

[2] e-Stat: http://www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do.

[3] UK data service: http://ukdataservice.ac.uk.

[4] Eurostat: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/.

[5] Data.gov: http://www.data.gov.

[6] Open data index: http://index.okfn.org.

[7] Open Knowledge Foundation: http://okfn.org.

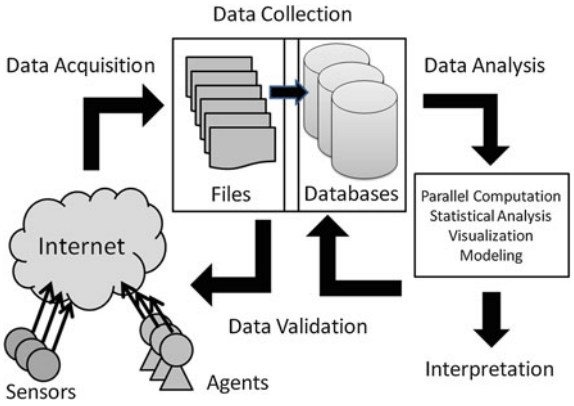


Fig. 1.7 A schematics illustration of the platform of data-centric sciences. Data are generated by sensor devices and both human and machine agents in actual environments. These data are transmitted through network infrastructure to a data platform. The data are collected as files in computers and stored in databases. Both validation and verification of the data are needed in order to improve the data quality. Researchers and practitioners use the data for analysis and decision-making

In fact, these data are no answers, but inspire questions to researchers and practitioners. Data just tells us how a phenomenon behaves, and at the same time, it asks us why a phenomenon behaves as we observe. Data is a collection of observations or facts. We first need to extract common parameters (variables) from the data. After that, we have to find relationship among parameters and attempt to construct a model of the phenomenon. This is an important task in explanatory data analysis.

If we can also accumulate data from sensors, and these sensors are connected with the Internet, then we can additionally accumulate data on the state of our world [60]. These sensors are called sensor networks [207]. Research in the area of sensor networks has been active at several levels, starting from the component level, the system level, and up to the application level. This has already been set up for geophysics and climates, where many sensors automatically collect data on temperature, winds and earthquakes. For example, Japan National Research Institute for Earth Science and Disaster Prevention has a nationwide broadband seismograph network called F-net,[8] and the Japan Meteorological Agency provides weather, climate and earthquake information in Japan.[9] Incorporated research institutions for seismology gives earthquake information across the globe.[10]

The data are automatically collected and stored. Eventually, our smart phone and computers will play the role of sensors in our society. Blog data and twitters can be another of these. Usage of mobile phones can be used as a further measure of human activities [166].

[8] F-net: http://www.fnet.bosai.go.jp/top.php?LANG=en.

[9] http://www.jma.go.jp/jma/indexe.html.

[10] http://www.iris.edu/seismon/.

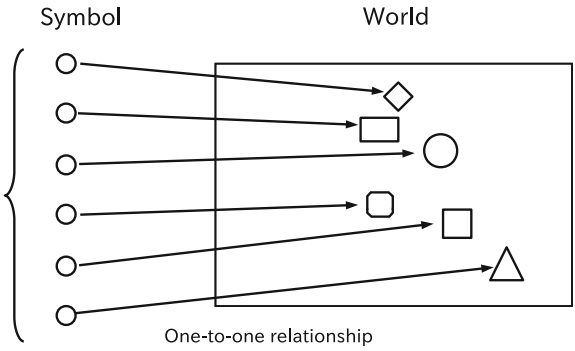


Fig. 1.8 Examples of one-to-one relationship. Each symbol has a one-to-one relation to an object or concept

There are five important steps of data-centric sciences: Data Acquisition, Data Collection, Data Validation, Data Analysis, and Interpretation. This is understood as a data analysis pipeline [37, 204].

Normally, data are noisy and context-based. We need to clean the data and store them as digital files (raw data or meta data). Several types of database servers can be used to handle and select data by using some conditions. After pre-processing, data are inserted into the database servers.

Some data are structured as Extensible Markup Language (XML). The Structured Query Language (SQL) standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allow customers to choose between vendors. XML databases have recently become available in both commercial and open usages. Some SQL databases include XML extensions or XML parsers. The XML extensions provide functions to insert and search data formatted as XML. The XML parsers provide the ability to read an XML file/string and extract its contents according to the structure. Several NoSQL databases or object-oriented database servers are enabling us to collect and search data with methods different from those of traditional relational databases.

In further advances, we also use data in order to control our environment. This is called a cyber-physical system (CPS) [129], which consists of distributed computation interconnected by computer networks that monitor and control switched physical entities interconnected by physical infrastructures [157, 158]. The data flow from an actual environment, computation based on the collected data, and control over our environment based on computations are main focuses [183].

1.3.2 Meaning of Data

Data is described as symbols associated with objects or concepts in the real world. Ideally, the relationship between a series of symbols and an object (or a concept) is expressed as a one-to-one relationship. Figure 1.8 shows examples of this. Mathematically, this type of relationship can be described as an injective function between a set of symbols and a set of objects. Let f be a function whose domain is

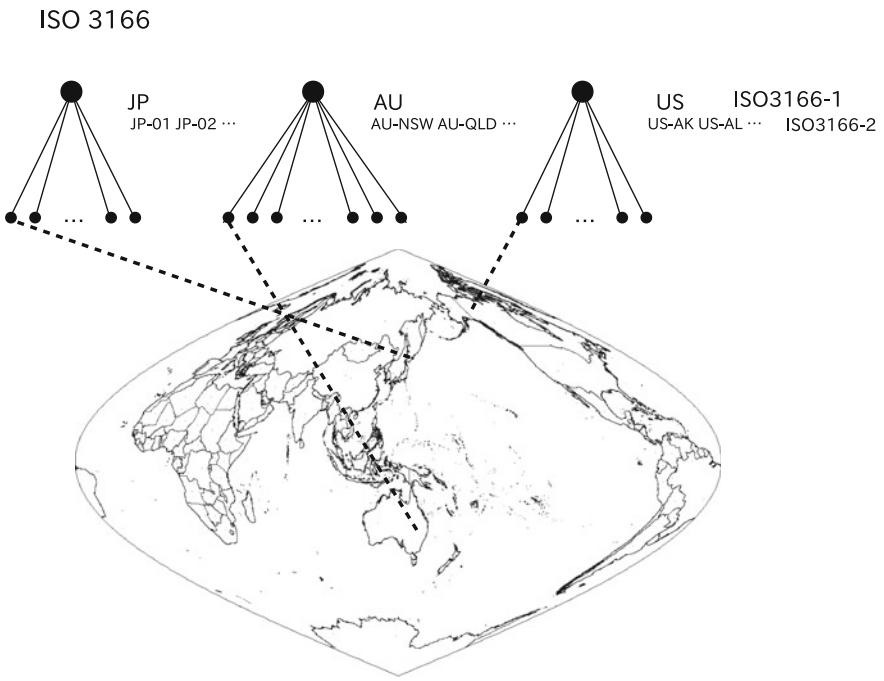


Fig. 1.9 ISO 3166: ISO 3166-1 provides country codes and ISO 3166-2 provides regional codes in each country

a set A. The function f is injective if

$$\forall a_1, a_2 \in A,\ f(a_1) = f(a_2) \Rightarrow a_1 = a_2, \tag{1.3}$$

which is logically equivalent to

$$\forall a_1, a_2 \in A, a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2). \tag{1.4}$$

Generally speaking, the injective function f is defined as a coding system (determined by researchers, practitioners, policy makers, and so on). Therefore, functions often depend on the coding systems. Namely, since symbols do not contain any relationships, the relationship between the symbols and entities or quantities should be imposed by the designers of datasets.

Thus, data is provided with an explanation of symbols. When we want to analyse data, we also carefully read the definition of a sequence of symbols describing the data and their explanation. This is defined as a database schema in a data management system in general. The formal definition of database schema is a set of formulae (sentences) called integrity constraints, which are imposed on a database. Normally, a conceptual schema provides meanings of symbols in the real world.

Table 1.2 Example of ISO 3166-1

Country name	Alpha-2	Alpha-3	Numeric
Australia	AU	AUS	036
Canada	CA	CAN	124
China	CN	CHN	156
Germany	DE	DEU	276
Japan	JP	JPN	392
Republic of Korea	KR	KOR	410
Netherlands	NL	NLD	528
Poland	PL	POL	616
Singapore	SG	SGP	702
Switzerland	CH	CHE	756
United Kingdom	GB	GBR	826
United States	US	USA	840

Three types of country codes are defined

To avoid confusion among different coding systems, standards of codes have already been defined. We share several types of codes defined by the International Standardisation Organisation (ISO). For example, we can denote countries, regions, currencies and so on identically with these. Figure 1.9 shows a conceptual illustration of the one-to-one relationship between codes of ISO 3166 and countries. ISO 3166-1 provides country codes and ISO 3166-2 provides regional codes in each country. There is hierarchical structure in ISO 3166. Tables 1.2 and 1.3 show examples of country codes (ISO 3166-1) and regional codes (ISO 3166-2). Even in ISO 3166-1, there are three types of country codes called Alpha-2, Alpha-3, and three-digit numeric. These codes should be used in order to allow every researcher to access, refer to and share the results of the data when we need to acquire, collect, analyse and publish data at an international level.

Some international organisations other than ISO also define their own international common codes. For example, the International Air Transportation Association (IATA) defines airline and airport codes. A 2-letter code of an airline and the 3-letter code of an airports are used in the IATA system, while, the International Civil Aviation Organisation (ICAO) issues the 3-letter code of an airline and the 4-letter code of an airport. Table 1.4 shows some examples of airport codes. Since IATA and ICAO independently issue their codes, three types of situations exists; (i) both IATA and ICAO issue an airport code, (ii) IATA issues an airport code, but ICAO does not issue one, and (iii) IATA does not issue an airport code, but IATA issues one. In the case of (i), we can identify airports by using both IATA and ICAO codes. In this case, an exchange table between IATA and ICAO codes are useful. In the cases of (ii) and (iii), we can only determine an airport by using IATA or ICAO.

In terms of time, there is Coordinated Universal Time (UTC), which is the primary time standard by which the world regulates clocks and time. The current version of UTC is defined by International Telecommunications Union Recommendation (ITU-R TF. 460-6). We can express time zones around the world as positive or negative

Table 1.3 Examples of ISO 3166-2

Prefecture name	ISO 3166-2:JP
Hokkaido	JP-01
Aomori	JP-02
Iwate	JP-03
Miyagi	JP-04
State name	**ISO 3166-2:AU**
New South Wales	AU-NSW
Queensland	AU-QLD
South Australia	AU-SA
Tasmania	AU-TAS

The regional codes are defined as sub-codes of ISO 3166-1

Table 1.4 Some examples of airport abbreviation and name

IATA	ICAO	Airport name	Country code	Regional code
PEK	ZBAA	Beijing Capital International Airport	CN	CN-11
CBR	YSCB	Canberra International Airport	AU	AU-ACT
DRS	EDDC	Dresden Airport	DE	DE-SN
FRA	EDDF	Frankfurt am Main International Airport	DE	DE-HE
GVA	LSGG	Geneva Cointrin International Airport	CH	CH-GE
ICN	RKSI	Incheon International Airport	KR	KR-28
KIX	RJBB	Kansai International Airport	JP	JP-27
AMS	EHAM	Amsterdam Airport Schiphol	NL	NL-NH
LHR	EGLL	London Heathrow Airport	GB	GB-ENG
YUL	CYUL	Montreal/Pierre Elliott Trudeau International Airport	CA	CA-QC
NRT	RJAA	Narita International Airport	JP	JP-12
PHX	KPHX	Phoenix Sky Harbor International Airport	US	US-AZ
SFO	KSFO	San Francisco International Airport	US	US-CA
SHA	ZSSS	Shanghai Hongqiao International Airport	CN	CN-31
SIN	WSSS	Singapore Changi International Airport	SG	SG-04
WAW	EPWA	Warsaw Chopin Airport	PL	PL-MZ
ZRH	LSZH	Zurich Airport	CH	CH-ZH

offsets from UTC. Table 1.5 shows examples of time zone acronyms and offset. For example, Japan Standard Time (JST) has +9:00 as its offset, and Central Standard Time in United States (CST) shows −5:00 as its offset.

There is a lifetime of data in these standard codes. We need to mention updates of these international codes. Some of them change or are newly inserted and deleted. This is one of hard problems when we analyse data with standard codes.

Table 1.5 Time zone acronyms and offsets from UTC

Acronym	Time zone	Offset
AET	Australia Eastern Time	UTC+10:00
JST	Japan Standard Time	UTC+9:00
CTT	China Taiwan Time	UTC+8:00
VST	Vietnam Standard Time	UTC+7:00
IST	India Standard Time	UTC+5:30
EAT	Eastern African Time	UTC+3:00
ECT	European Central Time	UTC+1:00
UTC	Universal Coordinated Time	UTC
CAT	Central African Time	UTC−1:00
UTC	Greenwich Mean Time	UTC
BET	Brazil Eastern Time	UTC−3:00
CNT	Canada Newfoundland Time	UTC−3:30
EST	Eastern Standard Time	UTC−5:00
CST	Central Standard Time	UTC−6:00
PNT	Phoenix Standard Time	UTC−7:00
AST	Alaska Standard Time	UTC−9:00
HST	Hawaii Standard Time	UTC−10:00

1.3.3 Understanding the World from Data

How can we understand the world from data? We have two types of understanding:

- qualitative understanding
- quantitative understanding

If we can change our behaviour through decision-making from the results of investigation, then our investigation turns out to be meaningful. If not, it is meaningless.

Roughly speaking, we often use four types of descriptions of data:

- time series
- spatial description
- network representation (relationship)
- text

To analyse these data, we may employ:

- time series analysis
- spatial analysis
- network analysis
- natural language processing

Of course, we can use combination of these methods in data analysis.

Time series analysis uses several types of methods to estimate parameters of models. Regression analyses are the most popular methods to quantify relationships

within data. Explained and explanatory variables can be assumed to determine causality (cause and effect). Autoregressive models are used to determine the effects of past observations on a present observation.

Every variable observed in our society is non-stationary. We need to carefully consider the non-stationary nature of the temporal development of human behaviour. In order to analyse a non-stationary time series, we need to have a model of non-stationarity. One of the models for non-stationary time series is some combination of locally stationary time series. Non-stationary time series are assumed to consist of several locally stationary segments. For this purpose, segmentation procedures or segmented regression analyses have been developed in mathematical economics or statistics. Regime switching models and a GARCH process are used in econometric and finance. We will treat these topics in Chaps. 3 and 6.

Moreover, network structure of socioeconomic-technological system is also modelled as a network. There are several methods to quantify states of network structure in network analysis. A degree distribution, average path length, betweenness centrality and network entropy are often used. We will treat these topics in Chaps. 3 and 4.

Spatial analysis is often used in order to quantify the spatial distribution of data. Therefore, the population density and geographical positions of humans are some of the most important issues of social sciences. We will treat spatial data in Chaps. 3, 4, 7 and 8.

According to Goodchild [74], every human is able to act as an intelligent sensor, and in that sense, the earth's surface is currently occupied by more than seven billion sensors. In this way, urban population centres of humans have been successively studied by many researchers [20, 22, 62, 73, 123, 130, 138].

Time series with spatial information have been studied in various ways. The mobility of individuals can be currently studied from geographical information. We have already obtained data on population density in real-time. Reades et al. [24, 147] show that real-time demographic data can be collected by using logs of mobile phones [166]. The spatial patterns of human communication and their mobility can be modelled in spatio-temporal data. We will treat these topics in Chaps. 3, 4, 7 and 8.

In addition, human communication can be described as symbols. These symbols are formed into languages. Natural language processing (NLP) provides methods that enable computers to derive meaning from human or natural language input. NLP tasks include automatic symmetrisation, discourse analysis, machine translation, morphological segmentation and named entity recognition, information retrieval (IR) and information extraction (IE).

For example, the automatic symmetrisation is to produce a readable summary from a chunk of text. The discourse analysis is a work to identifying the discourse structure of connected text. Recently, Evans and Foster propose metaknowledge, which is defined as knowledge about knowledge [59]. The growth of electronic publication and informatics archives makes it possible to harvest vast quantities of the metaknowledge. Metaknowledge research also investigates the effect of knowledge context on content. The metaknowledge research requires methods of NLP and statistics.

1.4 Concept

1.4.1 Why Do We Measure?

Galileo Galilei, who was a famous Italian natural philosopher, astronomer and mathematician, told us to measure what is measurable and make measurable what is not so measured. In fact, his interests were mainly in natural sciences. However, his saying is also alive in the goals of data-centric social sciences. We can enthusiastically measure our environment and socioeconomic-technological systems using data, but what is the goal of measurement in this research? Measurement is not a single goal but a part of the process of improving something. We cannot control what we cannot measure. Measuring what we have an interest in and obtaining insights from observation enable us to improve something. Galileo's quote leads us to the process for improvement. The purpose of data analysis can be to extract insights to determine our actions in real environments.

This forms a hierarchical structure from data to knowledge. We extract information from data and construct partial knowledge from some segments of information. We always desire to obtain knowledge since we want to behave intelligently in the real world. The final goal of our data analysis in the real world is not to just obtain knowledge but to determine how best to act in real environments.

The classical definition of knowledge by Plato specifies that a statement must meet three criteria in order to be considered knowledge: it must be justified, true, and believed. This means that we need to have some experience to confirm its truth in our world in order to construct our knowledge. From the scientific point of view, knowledge is justified through the scientific method, consisting of the collection of data through observations and experiments, and the formulation and testing of hypotheses.

Sir Francis Bacon established and popularised an inductive methodology for scientific inquiries. According to his definition of the scientific method, science needs to collect evidence that is both observable and reproducible—and apply concrete inference rules to it. In this case, it is important to collect data from experiments and observations, hypothesise, and falsify hypotheses based on the data. The scientific method consists of four perspectives; experiments, theory, computation and data. The data-centric social sciences appear to focus on characteristics of data and its properties.

Given the problem of induction, a finite number of confirming observations cannot verify a universal generalisation. Sir Karl Popper proposes falsification as a solution to the problem of induction. He also stresses the problem of demarcation, distinguishing the scientific from the unscientific, and makes falsifiability the demarcation criterion.

Desiring to approach wisdom, we carry out scientific research with scientific methodology. Data-centric social sciences should also be science, based on actual data on socioeconomic-technological systems. Therefore, data-centric social sciences utilise methodologies of data-centric science to deal with societal issues. Data-

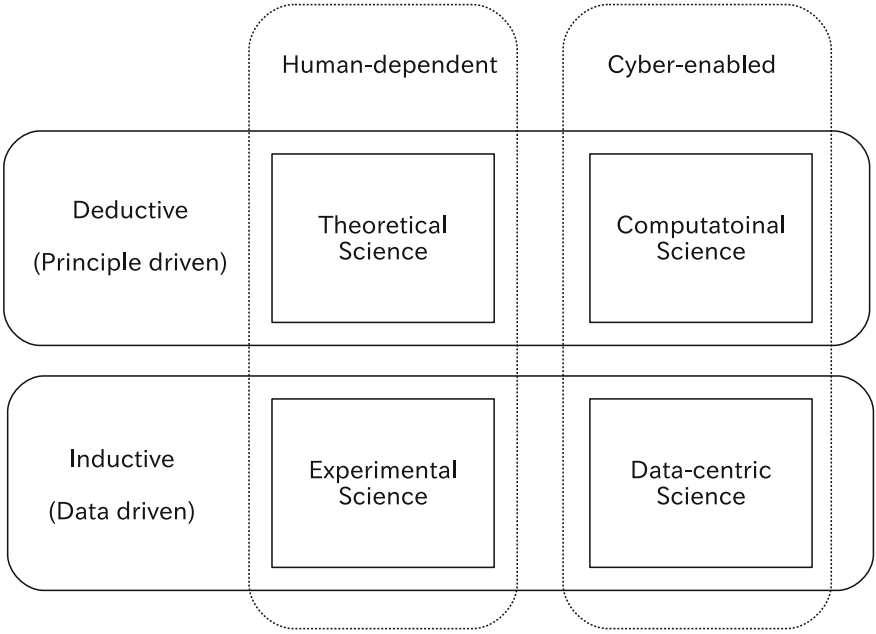


Fig. 1.10 Four methodologies for scientific work

centric science is a terminology to show a new scientific concept based on data, which has been used by some scientists since 2007. According to Kitagawa [118], computational science and data-centric science are newly established cyber-enabled deductive and inductive methods. He proposes four faces constructed from inductive and deductive methods, and human-dependent and cyber-enabled methods. Figure 1.10 shows the four methodologies that drive scientific researches. In this pictorial illustration, data-centric science is positioned as a fourth science with experimental science, theoretical science and computational science.

In inductivism, in order to construct a model, we assume that a collection of data is generated from the same mechanism. This is called *principle of the uniformity of nature*. The principle of the uniformity of nature appears to be true; however, principle of uniformity of society can not be validated. Since our humans have free will, individual behaviour does not always show the same response to the same stimuli or perception. In fact, human beings often show the same response to stimuli with high probability, if there is a strong condition. This situation is sometimes observed. Only in this case, the principle of uniformity in society seems to be true. In inductivism, in order to find a law or pattern, we need to assume the principle of uniformity in observations. Since data-centric science is inductive, we must carefully consider the principle of uniformity in society. Statistical regularity is often observed in socioeconomic-technological system. The statistical regularity implies that random events exhibit regularity when repeated enough times or that enough

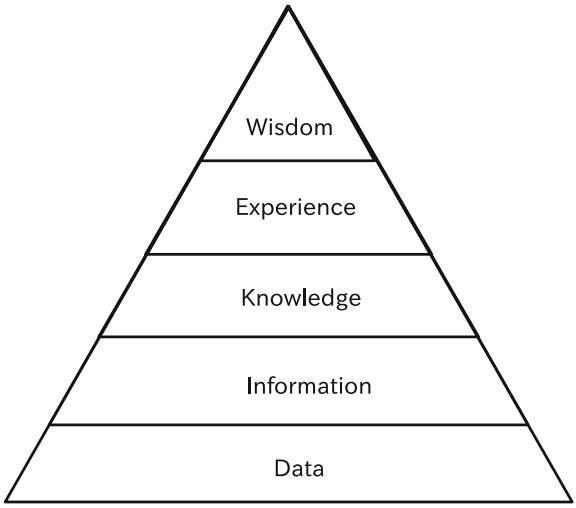


Fig. 1.11 A hierarchical structure from data to wisdom

sufficiently similar random events exhibit regularity. This covers the law of large numbers, central limit theorems and ergodic theorems. If we observe random events with different regularity repeatedly, then we need to separate them into each regularity. Finite mixture models can deal with a mixture of several types of different reasons.

According to a qualitative definition of information by Gregory Bateson, information is defined as "a difference which makes a difference". In his definition, a message and an observing system are assumed. The inner state of the observing system changes if the observing system recognises a meaning in the message. This can be seen as that what is included in the message makes the inner state of the observing system change. Something in the message is recognised as information. For example, if you see clouds in the morning, and go out with an umbrella. Under this situation, the "clouds" provide you with information. Namely, you recognise a meaning in the clouds. Otherwise, the "clouds" provide you with no information and you do not recognise any meanings in the clouds.

I would like to close this section with the following conclusion: there is a hierarchy structure from data to wisdom. The "data" collected from observations and experiments leads to "information". Several segments of "information" create "knowledge". The "knowledge" is confirmed in the real world through "experience". This "experience" allows us to determine our best actions in the world with "wisdom".

Figure 1.11 shows a conceptual illustration of a hierarchical structure from data to wisdom. As mentioned in Sect.1.3.2, the data is the closest to real environments. Data consist of records described by symbols generated through observations in the real world. Information is extracted from the data, and it is used to construct knowledge. To create wisdom from segments of knowledge, we need to justify the knowledge and confirm that it is true in real environments. This is a reconstruction mechanism

from data to wisdom in an inductive inference (to derive a general pattern or rule from every event and object). Final justification appears to be done in a deductive inference (to infer each event or object from a fundamental principle).

1.4.2 What is a Model?

1.4.2.1 The Purpose of Models

The purpose of modelling is not unique. We construct models for several purposes. Joshua Epstein proposes the purposes of developing models are to [58]:

- prediction
- explain
- illuminate core dynamics
- suggest dynamical analogies
- discover new questions
- promote a scientific habit of mind
- bind outcomes to plausible ranges
- illuminate core uncertainties
- offer crisis options in near-real time
- demonstrate trade-offs/suggest efficiencies
- challenge the robustness of prevailing theory through perturbations
- expose prevailing wisdom as incompatible with available data
- train practitioners
- discipline the policy dialogue
- educate the general public
- reveal the apparently simple (complex) to be complex (simple)

From a data-centric point of view, models are used with a different purpose in each step:

- discovering questions and aims in a research design step
- a guide of data collection in a data acquisition step
- explanation, prediction and inference in an analysis step
- decision support in an application phase

1.4.2.2 Unknown Unknowns

In order to understand the relationship between data and models, I introduce here the concept of the Johari House with four rooms [85]. This is constructed from two dimensions of Self and Others. Each dimension has two types: known and unknown. Table 1.6 shows the concept of the Johari House. The first room is called “Area”. The Area is the part of ourselves that we see and others see. The second room is

Table 1.6 Johari's house with four rooms

	Known to others	Unknown to others
Known to self	Area	Façde
Unknown to self	Blind spot	Unknown

Table 1.7 Four faces consisting of model and data

	Model is known	Model is unknown
Data is known	KK	KU
Data is unknown	UK	UU

called "Blind Spot". This is the aspect that others see but we are not aware of. The third room is called "Façde". This is our private space, which we know but keep from others. "Unknown" is the most mysterious room in that our unconscious or subconscious part is seen by neither ourselves nor others.

Suppose that we ourselves correspond to the data, and that others correspond to models. Using the Johari House, we can construct four faces consisting of two dimensional concepts of Model and Data. Table 1.7 shows four rooms; Known Known (KK), Known Unknown (KU), Unknown Known (UK), and Unknown Unknown (UU).

(KK) Known Known (Model and Data): This is observed in problems where we have a well-tested model and much data. The model is used to forecast future situations and/or infer unobservable portions from the data.
(KU) Known Unknown (No Model and Data): This is the most commonly encountered situation, specifically for problems in the social and behavioural domains. We have a lot of measured data, and want to construct a model that fits that data.
(UK) Unknown Known (Model but No Data): The notion of "knowing" what is unknown seems to be paradoxical. How can we know the unknown? The degree of this paradox is related to the distinction between implicit and explicit knowledge. This is deeply related to a transition process from the implicit to the explicit and emergent properties. We can assume or infer that there is a mechanism of a phenomenon, however, we do not have a concrete way to observe the phenomenon. In this case, we face this situation. What type of methodology do we have available to bring out these emergent properties?
(UU) Unknown Unknown (No Model and No Data): It is very difficult to imagine what kinds of events might be included here. How can we imagine an unknown that is definitely unknown? One possible answer is related to our lifetime. If we think about types of events that actually have happened and then consider variants of those events that have never yet been seen, we start edging into the domain of unknown unknown. In general, we cannot know what we have never seen.

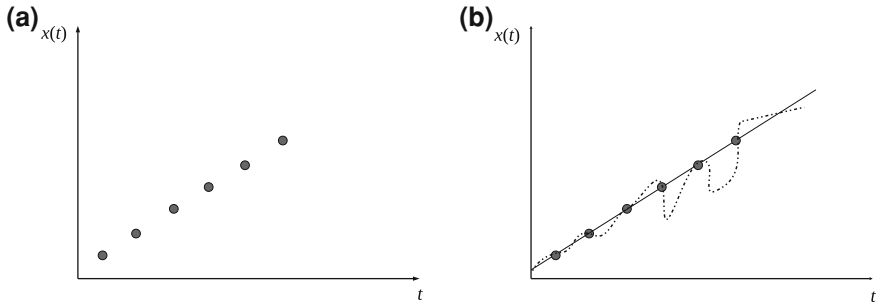


Fig. 1.12 **a** How do you draw a line to fit several points? **b** A line fitting to the points is a possible answer, but it is limited to the line

Specifically, the Unknown Unknown is often mentioned in the context of extreme events. The extreme events are highly improbable events which provide the significant impact. These are often observed as phenomena in natural disasters and synchronous behaviour in markets and society. They are treated by using extreme value theorem or some power-law distributions mathematically [2, 56, 57]. The probability distribution of impacts is well-fitted with a power law distribution. We will treat the method to estimate risks of the power law distribution in Chap. 5.

Nassim Nicholas Taleb calls the extreme events Black Swan events in his book [195]. A black swan event is positive or negative and an event that is deemed improbable yet causes massive consequences.

Thus, in order to investigate Unknown Unknown, we need to understand data of our society from a historical perspective. How do we infer the Unknown Unknown in our society? We can deal with these rare events: by using catalogued data of specific events described in historical documents and synthesising data from different sources, we can approach to Unknown Unknown.

1.4.3 Models for Socioeconomic-Technological Systems

We want to examine models of socioeconomic-technological systems in this section. Models of social systems appear to have many variables. Currently, our society consists of many individuals who possess properties (knowledge, relationship, goods, information and preferences). To describe our society in more detail, we have a tendency to need more variables. Some of them are hierarchically structured. The bottom-level models are used in the top-level models.

1.4.3.1 Agent-Based Models

Axelrod has suggested that agent-based models can be useful tools for "thought experiments" and clarification of theory. There are many types of agent-based models used in various problems of social sciences [11, 89, 128, 181, 192, 199]. LeBaron proposes agent-based computational finance [128]. Tesfatsion considers agent-based computational economics [199, 200]. Helbing [89] classifies two types of agent-based models (detailed models and simple models) and analyses their advantages and disadvantages.

A detailed agent-based model attempts to describe the world with agents whose behaviour is modelled with many parameters. This type of model can best be described by using object-oriented languages. An object-oriented language defines a class having inner variables and input and output interfaces. An object is produced from a class (instance). Objects can also interact with each other. We can use other objects in a class as well as variables. This feature can be used to describe agent-based models. Agents are modelled by using a class having some inner variables, input and output interfaces. Interactions among agents are also modelled as a class having several objects of agents. Based on this philosophy, major software platforms (NetLogo, MASON, Repast, Java Swarm and Swarm) are developed for agent-based simulation [165]. SOARS [197] is also an agent-based simulation language designed to describe agent activities according to roles within social and organisational structures. The U-mart[11] is a simulator for artificial market [182], allowing both human traders and software agents to participate in an artificial market to trade securities via the Internet.

In the context of computational social science, models that grasp as many details as possible have been developed. In fact, these try to implement many features of socioeconomic systems, but there are some problems with this approach. (1) It is difficult to specify parameters of detailed models in a real context. Models with many parameters have a large variety of different solutions. Imagine that we define an agent with 10 variables. Suppose that 10 small groups consisting of 10 agents are connected through some agents. This model has $10 \times 10 \times 10 = 1,000$ parameters at least. (2) It is hard to calibrate parameters of detailed models because the number of parameters is large. (3) It is not easy to determine the goodness-of-fit of detailed models with many parameters. (4) It is not easy to interpret the underlying mechanism of a certain phenomenon. Nevertheless, Yang et al. [223] propose a genetic algorithm based on an inverse technique of an agent-based model that is applicable to fit simulation outputs with real data.

A simple agent-based model attempts to avoid complexity. It tries to express the models as simply as possible. A simple model aims at extracting some simplified and abstracted essences of the system. This approach is based on the "Keep It Simple and Straightforward" (KISS) principle for building a model. This principle is also known as Occam's razor or as the principle of parsimony. Albert Einstein once asserted, 'Make everything as simple as possible'. A simple model is expected to give a

[11] http://www.u-mart.org/html.

better understanding than detailed agent-based models. The model is used to guide researchers and provide a working hypothesis. This can include a simple ecological model as a socioeconomic model such as Ising-models [34, 99], Granovetter-type threshold models [79, 216] and voter models [39, 63, 141, 187].

There are some criticisms of simple models: They are too simple to use in decision-making. A simple model tends to be oversimplified and ignore other important parts of phenomena. To improve a simple model is to add elements, which implies that the successive improvement of the model includes making the model complicated. Improving a simple model repeatedly turns out to make it a detailed model.

1.4.3.2 Statistical Model

A statistical model assumes a true statistical distribution with parameters, and the data is assumed to be sampled within the statistical distribution. A statistical model is rather simple, but the reason for why the distribution is adequate is not necessarily specified. This is derived from stochastic processes and/or dynamical models. The model parameters are estimated by using the maximum likelihood procedure. The models are evaluated based on an information criterion, and an adequate model should be selected by maximising the information criterion. Several types of information criteria have been proposed [120]. The Akaike Information criterion is one of them. The Bayesian information criterion is also often used. The information criterion is defined from the maximum log-likelihood value and a penalty function in terms of the number of parameters and/or the number of observations (see Sect. 3.1.8). The created model can sometimes be derived from a simple model in a theoretical way or from a detailed model in a numerical way, but it is often empirically confirmed.

Data assimilation is the process by which observations are incorporated into a computer model of a real system [31, 148]. There are two purposes of the data assimilation. One is to improve accuracy and ability of numerical simulation using the actual data. In numerical simulation, we always need to determine adequate initial states and boundary conditions as well as parameters. If we have actual data, then we may estimate them by minimising the difference between observations and results obtained from the numerical simulation. The other purpose is to supplement missing data with numerical results obtained from the calibrated model and to modify observation errors. It is not practically impossible to obtain the observations at points homogeneously distributed in time and space. The data assimilation enables us to generate the data inferred at the points homogeneously distributed in time and place.

The data assimilation attempts to compute best estimates of probability distributions to explain the data. Recursive Bayesian estimation is an ideal method; however, probabilistic analysis is usually simplified into a computationally feasible form. In general, the probability density function (PDF) in time can be described by the Fokker–Planck equation [169] (we will employ a one-dimensional Fokker–Planck equation in Sect. 3.1.12).In the case of a continuous stochastic dynamical system with partial observations distributed at discrete times, an ideal data assimilation scheme would be given by solving the Fokker–Planck equation for the time interval between

observations. The scheme should be implemented as a rule that modifies the PDF in both time and space by using all the available observations.

The fundamental difficulty of this approach, however, underlies the high dimensionality of the state space. Moreover, it is impossible in practice to obtain the initial probability density function. As the method of the Fokker–Planck equation requires computational resources, various approximations operating on simplified representation of the probability distribution are used [31]. If the probability density function is normal and the linear dynamics and linear relationship between observations on the system's state variables are assumed, then the probability density function can be characterised by its first and second moments. Namely, it can be expressed by its mean and variance-covariance matrix, which is called the Kalman filter (KF) equations [108, 118]. In the case of a large number of degrees of freedom in the state, we also need to use approximations.

To form the data assimilation, we need the four elements:

- numerical simulation
- observation data
- statistical methods
- high performance computing (HPC)

The data assimilation requires both a model and data in numerical simulation. There are several methods of data assimilation. The fundamental underlying idea behind data assimilation is to reduce the difference between observations and the results obtained from the numerical simulation.

Suppose that the model of numerical simulation can be described as

$$\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_{t-1}, t) + \boldsymbol{\xi}_t, \tag{1.5}$$

where $\boldsymbol{x}_t$ is a column vector representing states, t is discrete time, $\boldsymbol{f}$ the assumed model and $\boldsymbol{\xi}_t$ a column vector expressing random disturbance. Consider the difference between the k-dimensional observations $\boldsymbol{y}_t$ and the k-dimensional results obtained from numerical simulation $\boldsymbol{f}(\boldsymbol{x}_t; t)$. Then, how do we find adequate results from numerical simulation minimising the difference between the observations and the results?

The simplest way is to use the least squared regression (see Sect. 3.1.9). Namely, the results of numerical simulation $\boldsymbol{x}_t$ is determined by minimising the squared error $||\boldsymbol{y}_t - \boldsymbol{f}(\boldsymbol{x}_t; t)||^2$. Normally, we need to seek values $\boldsymbol{x}_t$ around assumed values $\boldsymbol{x}_{t,b}$. Therefore, the evaluation function of this problem is assumed to be $||\boldsymbol{x}_t - \boldsymbol{x}_{t,b}||^2 + ||\boldsymbol{y}_t - \boldsymbol{f}(\boldsymbol{x}_t; t)||^2$. The linear model $\boldsymbol{f}(\boldsymbol{x}_t, t) = \boldsymbol{H}_t \boldsymbol{x}_t$ with a time-dependent matrix $\boldsymbol{H}_t$ is often assumed.

A more sophisticated way is to use Bayesian inference. Bayes' theorem is formalised as

$$p(\boldsymbol{x}_t|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t)}{\int p(\boldsymbol{y}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t}. \tag{1.6}$$

The PDF $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$ on the left hand side describes a conditional probability distribution of the state variables under the observations $\boldsymbol{y}_t$. If we assume that $p(\boldsymbol{x}_t)$ and $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ are normal

$$p(\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{V}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{x}_{t,b})^T \boldsymbol{V}^{-1}(\boldsymbol{x}_t - \boldsymbol{x}_{t,b})\Big), \tag{1.7}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{R}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{H}_t\boldsymbol{x}_t)^T \boldsymbol{R}^{-1}(\boldsymbol{y}_t - \boldsymbol{H}_t\boldsymbol{x}_t)\Big), \tag{1.8}$$

and the dynamics $\boldsymbol{f}$ is linear ($\boldsymbol{f}(\boldsymbol{x}_t, t) = \boldsymbol{H}_t\boldsymbol{x}_t$), then we can estimate the state variables $\boldsymbol{x}_t$ by using

$$\boldsymbol{x}_{t,est} = \boldsymbol{x}_{t,b} + \boldsymbol{V}\boldsymbol{H}_t^T(\boldsymbol{H}_t\boldsymbol{V}\boldsymbol{H}_t^T + \boldsymbol{R})^{-1}(\boldsymbol{y}_t - \boldsymbol{H}_t\boldsymbol{x}_{t,b}).$$

This is referred to as optimal interpolation (OI) [44, 45]. The third method is a KF algorithm [108]. The KF equations describe the time evolution of both the mean and the covariance. According to Carrassi [31], in the case of linear dynamics, a linear observation operator and observational and system noise that are both Gaussian, white in time and mutually uncorrelated, the KF equations give the optimal linear estimate of the state of the system by propagating the associated error covariances, along with the state estimates. In the nonlinear case, the PDF cannot be described by a finite set of parameters, and the extended Kalman filter algorithm is required by extending the linear results to the nonlinear case.

1.4.4 Forecasting the World from Data

Suppose that several points $x(t)$ in terms of t, which are assumed to be observations, are plotted on a graph as shown in Fig. 1.12a. If you have to draw a fitting curve, how do you draw a line? You may want to draw a solid line as shown in Fig. 1.12b. This is a typical example showing a prediction problem from past observations. In fact, you can draw a line passing through the points. However, this line is not always a good model of these observations.

If several types of functional forms are permitted, then we may obtain other types of solutions. The possibilities are infinite. In our habitual way of thinking, we may prefer a line, but this is just a belief. If we have a nonlinear model in our brain, we can find other types of solutions.

This problem is referred to as the problem of induction. David Hume worked on an explanation of how we are able to make inductive inferences. According to Hume, we tend to believe that things show regular patterns, and that the behaviour of objects will persist into the future, and throughout the unobserved present. This persistence of regularities is called uniformitarianism or the principle of the uniformity of nature.

The principle of uniformity of nature is the assumption that the same laws and processes that have operated in the past can be used to explain observations. Inductive reasoning always makes this assumption. This implies that we always assume a model when we make a prediction or reasoning. We want to believe that what we have experienced can be applied to our current observations. Further, we want to find a pattern from the data under some assumptions from models that we create during our experiences. We have a tendency to assume that what we will observe is similar to what we have experienced in the past.

However, Hume argues that we cannot rationally justify the principle of the uniformity of nature. Both demonstrative reasoning and probable reasoning are inadequate. If the assumption in our model fits with the data generating mechanism, then our prediction seems to work well. This is true, but this is not rationally justified.

Nelson Goodman also dealt with this problem as the new riddle of induction [75]. This problem is also mentioned in the grue paradox. Normally what we want to recognise often possesses more than two properties at the same time. This paradox is deeply related to the principle of the uniformity of nature and originates from the fact that we often use projectable predicates to discriminate something included in more than two groups at the same time. This situation is often seen in social sciences.

Moreover, individual behaviour may be affected by announcing prediction. Thus, the principle of the uniformity in society may be violated by reactions to this prediction. This feedback mechanism between observed systems and observing systems creates complexity in our society.

1.4.5 Designing the World from Data

In fact, the predictability of our society from the data turns out to be weak conclusions. However, we do have an ability to create the future of our society.

Figure 1.13 illustrates coherent multilevel theories to support policies on social design. At the micro-level, the dynamics emerge from the interactions of individuals. An agent-based model on societal issues is a favoured scientific way of investigating micro-level behaviour in agent-based simulations. Members of the micro-level change their awareness and decisions under the meso-status. A large number of the micro-level members and macro-level factors affect the meso-status.

Terano et al. [198] propose multilevel agent-based computer simulations. Kasuga et al. [110] propose a new technique of constructing the simulation model that can take decision making of the patient, the fire department and the medical institute into consideration by using an agent-based model approach.

These simulations do not generate predictions of future system behaviour but give insights on possible system behaviour. This means that agent-based models are useful in the design of socioeconomic-technological systems, not from an empirical science point of view but from a normative science point of view.

The mesoscopic level can be described by stochastic dynamics. Financial time series are good examples of stochastic dynamics and can often be modelled by

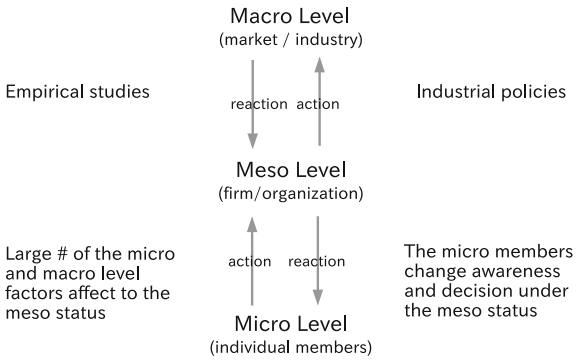


Fig. 1.13 A schematic illustration of multilevel systems. micro-level to macro-level through mesoscopic description

stochastic differential equations. The macroscopic level is institutional or national, which seems to show more dynamical behaviour.

Social policy is inevitably expressed in natural language within a legal framework for implementation. Nevertheless, numerical simulation by using an agent-based model must be described by some algorithms or mathematics. Johnson [103] also proposes the relationship between policy informatics and Big data in the context of policy design in "systems of systems of systems".

Social informatics is not necessarily policy-driven and can be pure research. Currently, big data plays an important part in social informatics and the development of new models of social processes.

In policy design, three types of methods can be considered;

- evidence-based design (EBD)
- simulation based design
- evolutionary design

Evidence-based design originated in the health-care industry as a combination of evidence-based medicine and evidence-based practice. The evidence-based design is a process for creating or improving products buildings by using rigorous evidence to create benchmarks of current practice, achieve specific goals, and then monitor the success of the design to inform future decision making. Rules of policy can be formed based on evidence that we experienced. Simulation-based design is often used in engineering more recently. Agent-based simulations are sometimes used to design socioeconomic-technological system. Evolutionary design needs to be introduced to policy design for long term perspectives, since rules or laws of our society have been created through evolutionary processes. This view is referred to as "cultural evolution".

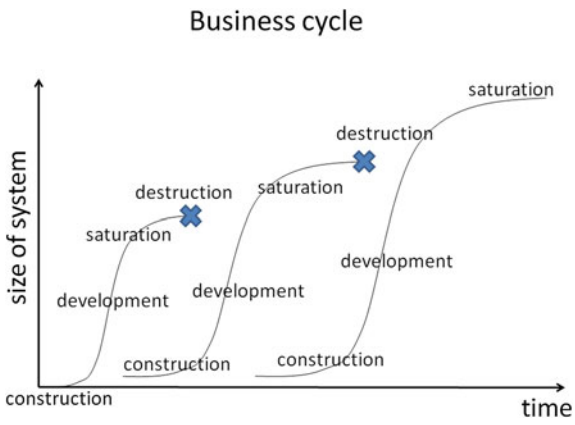


Fig. 1.14 Conceptual illustration of business cycles. A business cycle consists of four phases: construction, development, saturation, and destruction. New goods or services destroy a mature system reaching its saturation point

1.4.6 What is Collective Behaviour?

According to "The Theory of Economic Development" by Schumpeter [180], our modern capitalistic society creates business cycles, or boom-and-bust cycles. A cycle consists of expansion occurring at the same time in many economic activities. Business cycles may vary from more than one year to 10 or 12 years. Business cycles are inevitable under capitalism since these are associated with behaviour which firms employ to obtain gains from markets.

Business cycles have a long history of theoretical studies. Samuelson showed that a linear second-order ordinary differential equation based on multiplier and accelerator effects could generate these cycles in the GDP [172]. Hickes demonstrated that this type of cycle is sustainable by introducing a "ceiling" and a "floor" into his model [90]. Some researchers focused on non-linearity. Goodwin introduced a nonlinear accelerator in order to generate a sustainable cycle [76]. Kaldor captured the business cycle as a stable limit cycle [107].

The most influential theory among neoclassical economists is the real business cycle (RBC) theory [124]. The essential feature of RBC is to treat the impact of technological innovations as the most important cause of a business cycle. The market travels through its initial stage (construction stage), development stage, saturation stage, as shown in Fig. 1.14. Profits are generated only within the development stage. Eventually, the market becomes mature and the enterprises need to destruct the present market and create a new market. This movement generates a business cycle or boom-and-bust cycle.

This is an example of collective behaviour in human activity. Other types of collective behaviour are often seen in opinion formation. Opinions or preferences sometimes coincide due to social forces. There are several types of opinion formation models including the voter model, the heterogeneous voter model, the partisan voter model, opinion formation model on a network, the evolutionary model of language competition [1, 4, 35, 38, 39, 51, 61, 63, 92, 137, 140, 141, 150, 156, 187, 189,

205, 206, 226, 227]. The Bass model in marketing science explains how goods or services produced by new technology are accepted by users [16]. Jung [104] proposed a concept of synchronicity, which is the experience of two or more events as meaningfully related, even though these cannot be explained causally. Jung classifies synchronicity into three forms:

- The coincidence of a certain psychic content with a corresponding objective process which is perceived to take place simultaneously.
- The coincidence of a subjective psychic state with a phantasm (dream or vision) which later turns out to be a more or less faithful reflection of a synchronistic objective event that took place more or less simultaneously, but at a distance.
- The same, expect that the event perceived takes place in the future and is represented in the present only by a phantasm that corresponds to it.

1.4.6.1 Classical Models of Collective Behaviour

I next address some classical models of collective behaviour in this section. Collective behaviour or synchrony can be understood as the bifurcation of a dynamical system or the phase transition in an analogy to statistical physics. The simplest model of collective behaviour in statistical physics is the Ising model [99]. The Granovetter's type threshold model in sociology also shows similar properties to explain collective behaviour [79]. Schelling also provides one explanation for collective behaviour referred to as "contagion" [179]. Contagion of behaviour from one agent to other agents occurs when individual has an incentive to pay attention to the decisions of others. Kirman's model established "herding behaviour" (called conformity) in aggregate expectations stimulating agent interaction [117].

Ising Model

Although the Ising model [99] is a model of magnetism, it is often used as a model of opinion formation from a socioeconophysics approach [34]. Physicists have interests in the Ising model on several types of networks such as random graphs and scale-free networks [13, 21, 33, 84, 190].

I will introduce the one-dimensional Ising model based on Glauber dynamics [70]. Suppose that N particles interact in a random field where a state of the i-th particle flips between the values $\sigma_i = 1$ and $\sigma_i = -1$ randomly. For a sufficiently small period Δt, the individual probabilities $p(\pm 1, t)$ can be assumed as

$$p(\sigma_i, t) = \begin{cases} \dfrac{1}{1 + \exp(-I_i(t - \Delta t)/\theta)} & (\sigma_i = 1) \\ \dfrac{\exp(-I_i(t - \Delta t)/\theta)}{1 + \exp(-I_i(t - \Delta t)/\theta)} & (\sigma_i = -1) \end{cases}, \tag{1.9}$$

where $I_i(t)$ is an external field perceived by the i-th particle and $\theta(>0)$ represents a social temperature. For the sake of simplicity, we only assume an interaction of the one-dimensional case from the nearest neighbours

$$I_i(t) = \frac{J}{2}\big(\sigma_{i-1}(t) + \sigma_{i+1}(t)\big), \tag{1.10}$$

where J is a constant. If J is positive, then the state of the i-th particle takes the same direction as the nearest neighbours (a follower). Otherwise, it has a tendency to take a direction opposite to the nearest neighbours (a contrarian). We approximate the mean value as

$$\langle \sigma(t) \rangle = \frac{1}{N}\sum_{i=1}^{N} \sigma_i(t). \tag{1.11}$$

By using a mean field approximation that $\sigma_{i-1}(t)$, $\sigma_i(t)$ and $\sigma_{i+1}(t)$ are identical to $\langle \sigma(t) \rangle$, we have

$$\begin{aligned} \langle \sigma(t) \rangle &\approx +1 \times p(+1, t) + (-1) \times p(-1, t) \\ &= \frac{1 - \exp(-J\langle \sigma(t - \Delta t) \rangle/\theta)}{1 + \exp(-J\langle \sigma(t - \Delta t) \rangle/\theta)} \\ &= \tanh\Big(\frac{J\langle \sigma(t - \Delta t) \rangle}{\theta}\Big). \end{aligned} \tag{1.12}$$

By setting $\langle \sigma(t) \rangle$ and $\langle \sigma(t - \Delta t) \rangle$ into σ, a fixed point σ of Eq. (1.12) is derived from

$$\sigma = \tanh\Big(\frac{J\sigma}{\theta}\Big). \tag{1.13}$$

Here, we consider a fixed point of Eq. (1.12) in the case of $J > 0$. As shown in Fig. 1.15, we have two types of solutions; (a) $m = 0, \pm\sigma^*$ for $J/\theta > 1$ and (b) $m = 0$ for $J/\theta < 1$. Stability analysis of Eq. (1.12) tells us that $m = 0$ is an unstable fixed point and that $m = \pm\sigma^*$ are stable fixed points in the case of $J/\theta > 1$. In the case of $J/\theta < 1$, $m = 0$ is a stable fixed point. Therefore, for $J/\theta > 1$, all n particles converge to the same state: $\sigma_i = 1$ or $\sigma_i = -1$. This phenomenon is not confirmed for $J/\theta < 1$.

Collective behaviour is defined as the event where some particles take the same direction using an analogy of ferromagnetism. Since J represents the strength of positive interaction and θ represents the strength of disturbance, collective behaviour can be observed when positive interaction is stronger than disturbance.

Granovetter Threshold Model

Granovetter proposes a simple threshold model to explain collective behaviour in the case of binary decisions [79]. More recently, Watts and Dodds [216] studied influential hypothesis of public opinion formation. They introduced the threshold

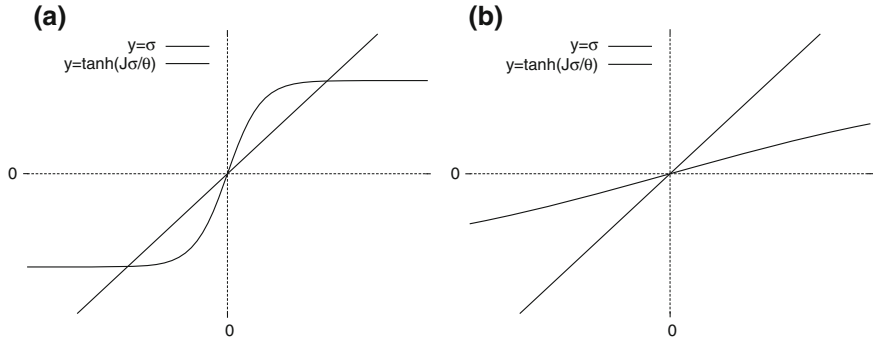


Fig. 1.15 Two types of solutions of the Ising model. **a** $J/\theta > 1$ and **b** $J/\theta < 1$

rule of decision-making on influence networks [215]. Their model proceeds from an initial state in which all N individuals are inactive (state 0), with the exception of a single, randomly chosen initiator i, who is activated (state 1) exogenously. Depending on the model parameters of i's neighbours, this initial activation may or may not trigger some additional endogenous activations. Subsequently, these newly activated neighbours may activate some of their own neighbours, who may, in turn, trigger more activations still, and so on, generating a sequence of activations, called "cascade" [215].

Granovetter suggests several examples of binary choice situations where threshold models could be applied;

- Diffusion of innovations
- Rumours and diseases
- Strikes
- Voting
- Educational attainment
- Leaving social occasions
- Migration
- Experimental social psychology

All the examples above can be expressed as a threshold model with the initial distribution of thresholds and the ultimate number or proportion of elements making each of the two decisions. We assume that there are two types of decisions: 1 and 2. A simple example is a case of a uniform distribution of threshold. Imagine 100 people have their thresholds, which are distributed as a uniform distribution. They are assumed to be able to see other decisions. There is one individual with threshold 0, one with threshold 1, one with threshold 2, and so on up to the last individual with threshold 99. This model can explain a "bandwagon" or "domino" effect. The person with threshold 0, the "instigator," must take decision 2. This activates the person with threshold 1; the activity of these two people then activates the person with threshold 2, and so on, until all 100 people join. In this case, the equilibrium

is 100. Next, we consider the case where we remove the individual with threshold 1 and replace him by one with threshold 2. In this case, the instigator takes decision 2 but other people do not take decision 2 since there is no person with threshold 1. Therefore, the collective behaviour does not occur.

The Granovetter threshold model is mathematically expressed as follows. Suppose that n people have their riot thresholds, which are distributed. The threshold x is sampled from the PDF $p(x)$ and the cumulative distribution function (CDF) by $\Pr[X \leq x]$ (see these definitions in Sect. 3.1.1). The CDF indicates the proportion of the population having thresholds less than or equal to x. Call the proportion of the population that has taken decision 2 by discrete time t $x(t)$. When knowing $x(t)$ for some t, we express the proportion at time $t + 1$ as a difference equation

$$x(t + 1) = \Pr[X \leq x(t)]. \tag{1.14}$$

Then, when the probability distribution of the threshold has a simple form, the difference equation can be solved explicitly to give an expression for $x(t)$ at any value of t. Thus, the equilibrium solution x may be solved as $x = \Pr[X \leq x]$ by setting $x(t + 1) = x(t) = x$. This implies that the shape of CDF and an initial condition determines the equilibrium state.

1.5 Literature Review

Recently, several researchers in a wide spectrum of fields have paid a remarkable amount of attention to massive amounts of comprehensive data. For example, search engines of web services need massive data about hyperlink connections among web pages, and electronic commerce systems need to cover information about various kinds of products. As a result of the development of ICT, the Advanced Information Society has already achieved a global profile, and it is gradually making our world smaller and smaller.

The term "information explosion" has been coined to describe this situation [54, 121]. This term refers to a situation in which the total amount of digital information created by individuals exceeds the individuals' information processing capability.

Researchers belonging to several different fields have the same interest in analysing data on socioeconomic-technological systems:

- statistics
- management sciences and marketing
- social network analysis
- socioeconophysics
- data engineering and computer sciences
- computational social science

Recently, several different disciplines have been attracted by rich data and often use a common framework of data-centric methodologies.

1.5.1 Statistics

Data scientists are industrial practitioners using data on socioeconomic-technological systems. They attempt to solve societal problems with knowledge obtained from data analysis, ICT and social implications [47]. These data scientists need to have knowledge of:

- mathematics, statistics, physics
- computer systems, ICT, software and hardware
- markets, specific fields of industrial sectors

Education for data scientists is now important issues in both academics and industries. Data scientists should be the people who understand how to find out answers to important business questions from current tsunamis of unstructured information. The roles of data scientists are akin to ones of quants of financial sectors in 1990's.

Business intelligence (BI) is an important field for organisations across all industries. A number of business sectors can continue to obtain benefits from the careful use of business intelligence. Business intelligence enables individuals to perceive information with little technical expertise. Contributions of business intelligence are divided into four categories: data presentation, creation of new knowledge, responsive and anticipate decisions, and improvement of planning for the future [171].

In many cases presence of wrong data is even worse than absence of the data, and it makes a harmful effect in decision-making and optimisation. Therefore, data validation is an important step [186]. There is further a problem that predicts future affairs from partially disclosed data [202].

Optimality in human society is not unique but this is derived from a trade-off relationship among sectors under multiple evaluation functions. Thus, this should be expressed as multi-objective optimisation problems. In the multi-objective optimality, the scalar concept of "optimality" does not apply directly. The concept of Pareto optimality should be introduced. The Pareto optimal for a multi-objective problem is that an N-dimensional vector $\boldsymbol{x}^* \in S$ exists if all other vector $\boldsymbol{x} \in S$ have a higher value for at least one of m objective functions $f_i(\boldsymbol{x})$, with $i = 1, \ldots, m$, or have the same value for all the objective functions.

To find the Pareto optimal solution cannot be computed efficiently in many cases. Even if it is theoretically possible, computationally they are reduced as a NP-hard problem. There are several methods to approximate the Pareto optima. To this context, evolutionary algorithms [101, 196] contribute to finding solutions.

1.5.2 Management Sciences and Marketing

Internet technologies have created new opportunities for companies to sell a variety of products and services via their e-commerce platforms. This has already mentioned by Arthur [8] in the 1990s. He stated that the mechanism that determines economic behaviour has shifted from the processing of resources and the application of raw materials to the processing of information and the application of ideas. Wymbs [219]

has suggested that the dramatic increase in the availability of information and a plethora of ways to manipulate it would both increase the number and diversity of new service businesses and cause the fundamental reconfiguration of existing service industries. He has proposed how uncertainty can be used by firms to create value by increasing their options in e-commerce. Wen et al. [218] also proposed strategies and models of e-commerce website design in the early stage of the Internet era.

These suggestions and proposals during the development of Internet have been implemented in current Internet environment. Recently, social media marketing, which is the process of gaining website traffic or attention through social media sites, is often used. There are several tools and methods [30] such as social media measurement, social network aggregation, social bookmarking, social analytics, opinion mining, sentiment analysis and sentic extraction.

Specifically, it is important to understand and manage electronic word of mouth (eWoM) in e-commerce. Kietzmann and Canhoto discuss classification of eWoM and implementation to manage it in social media marketing [114]. Li et al. [131] examine the factors related to eWoM that influence travelers' online hotel booking intention.

Techniques of social media measurement or social media monitoring have been developed [98]. Social media measurement is an activity to monitor information about firms or organisations from social media channels such as blogs, wikis, news sites, micro-blogs, video-photo-sharing websites and forums. After crawling digital files from these media channels, semantic analysis and NLP are applied to extract information. "Emotional polarity" is also an important measure in social media monitoring. The emotional polarity is a label indicating either positive, neutral or negative, assigned in each word of the lexicon. The emotional polarity technique allows us to characterise the emotional charge of words or sentences as either negative, neutral, or positive. Opinion mining focuses on opinion polarity detection, while sentiment analysis considers emotional influence. In general, existing approaches to opinion mining and sentiment analysis can be classified into three categories: keyword spotting, lexical affinity and statistical methods. Modelling emotional interactions in cyberspace and developing measures of social interactions are being studied by several researchers [42, 64, 66, 146, 222].

1.5.3 Social Network Analysis

Network analysis has often been used in the context of sociology [32, 127, 216] and economics [7, 9, 17]. Each kind of resource exchange is considered while constructing social network relationships among agents. These relationships maintained by agents are called "ties". The strength of a tie varies in time and ranges from weak to strong, depending on the quantity, quality and frequency of exchanges between agents [135]. Mark Granovetter proposes that weak ties are operationally strong for the diffusion of job information [78]. This is a very well-known example of social network analysis.

A long-standing traditional theory of public opinion and individual judgements has been affected by observations of social aggregations and mass behaviours. Formation and change of collective behaviour or public opinion are assumed to be a mechanism by which a certain action affects other future actions of agents. However, according to Olson [154], the logic of collective action is based on the assumption that individuals motivated by self-interest will avoid investing resources in a joint endeavour. This is a phenomenon known as "free riding". This means that agents' communication is very weak, and that common external stimuli drive their collective behaviour. Bad relationships sometimes spread in a group. Several researchers have had an interest in promoting and increasing good relationships. Oliver and Maxwell have emphasised the importance of the network of relationships in which interdependent agents are embedded [136, 153]. There are studies of network theory in several fields, such as social network analysis and data-driven network analysis.

Kim and Bearman [115] have developed an interesting model of opinion formation within a network. Agents increase their interest in participating in public processes if connected with others with higher interest levels who contribute, and they decrease their interest if connected to others with lower interest levels who defect. In this model, collective action occurs if and only if there is a positive correlation between interest and power/centrality. Therefore, heterogeneity of interests creates positive effects by "pulling up" a population's potential for participation.

Watts and Dodds [216] studied influential hypothesis of public opinion formation. They introduced the threshold rule of decision-making on influence networks [215] based on Granovetter threshold dynamics [79]. Their model proceeds from an initial state in which all N individuals are inactive (state 0), with the exception of a single, randomly chosen initiator i, who is activated (state 1) exogenously. Depending on the model parameters of i's neighbours, this initial activation may or may not trigger some additional endogenous activations. Subsequently, these newly activated neighbours may activate some of their own neighbours, who may, in turn, trigger more activations still, and so on, generating a sequence of activations called cascade [215]. Bahr and Passerini [15] have developed a statistical mechanics model of collective behaviour through an analogy to physical systems and studied under which conditions a group changes opinions and how this depends on the size of the group.

In the context of economics, banking sectors can be represented by using an agent-based model in a network. Lending and borrowing relationships among banks are expressed as a network [17]. Financial markets can be described by using agents, while the trading of stocks and currencies forms a bipartite network. To study the systemic risk of financial systems, Huang et al. [94] consider a bipartite banking network model composed of banks and bank assets and propose a cascading failure model to describe the risk propagation process. Companies have connections with other companies, and this can be expressed by using an agent-based model in a network [7]. The firm productivity is examined empirically from an international point of view. Mizuno et al. [143] estimate firm productivity for about 3.2 million firms from 30 countries.

Data-driven network analysis for socioeconomic-technological systems has been conducted by some researchers. Tomasello et al. [201] studied a large database of

publicly announced research and development (R&D) alliances. They have described the evolution of R&D networks in a large number of economic sectors from 1986 to 2009. They show that many properties of R&D networks are characterised by rise-and-fall dynamics, with a peak in the middle of the 1990s. This corresponds to IT and bio-tech doubles. Garas and Panos have studied the networks of collaboration between partners for projects carried out with the support of European Commission Framework Programmes FP5 and FP6. They have found that there is an increase in the average number of collaborative partners per institution when FP5 is compared to FP6, and that the number of signed contracts and the total number of unique partners has decreased [67].

1.5.4 Socioeconophysics

Socio-Econophysics is an interdisciplinary research field. Rich data enable physicists to investigate phenomena observed in our society from an empirical point of view. Many researchers belonging to the physics community have attempted to develop applications of methods in statistical physics in order to solve phenomena observed in economics and sociology [6, 7, 26, 36, 68, 134, 184, 188, 193, 194]. They often use methodologies in statistical physics, agent-based modelling and network analysis, which have been evolving over the past decade.

A large amount of data on financial markets has been available since electronic matching systems of financial markets have spread all over the world with the development of ICT. Computer trading can be done through electronic platforms, and settlement operations are done through electronic clearing systems. Real-time data of financial markets can be collected through direct market access (DMA) as well as historical data from centres of data providers. Some researchers have examined cross-correlations among various figures from the financial markets [53, 93, 133, 159, 176]. Podobnik et al. studied volume growth rates and volume changes for 14,981 daily records of the Standard and Poor's (S&P) 500 index over a 59-year period (1950–2009) [159]. Using detrended cross-correlation analysis, they found that there are power law cross-correlations between these indices. Bonanno et al. [23] studied correlation-based network analysis of financial equities. Researchers of econophysics focus on several types of scaling relationships observed in the financial markets in order to understand the behaviour of market participants [55, 139, 160, 213].

Studies of foreign exchange rates have been conducted by numerous researchers using various approaches based on statistical physics and time series analysis [3, 52, 82, 86, 97, 109, 132, 152, 167, 178]. Drożdż et al. have shown that exchange rate return fluctuations for main currency pairs are well described by non-extensive statistics and possess multifractal characteristics [52]. Rebitzky has studied the influence of macroeconomic news on exchange rates [167].

Kaltwasser [109] has further estimated the herding tendency in the foreign exchange market for three currency pairs, using the extended Alfrano–Lux model [3],

and has computed several unconditional moments of corresponding daily log-returns. Gworek et al. [82] have analysed the exchange rate returns of 38 currencies (including gold). They examined the cross-correlations between the returns of various currency pairs, as well as between their signs, and in this way, they have constructed a corresponding Minimal Spanning Tree for several base currencies. Liu et al. [132] applied cross-sample entropy (Cross-SampEn) to compare two different time series in order to asses their degree of asynchrony to the daily log-return time series of foreign exchange rates in the currency markets. Several types of models of financial markets are studied: agent-based models and stochastic processes are often used [149, 163, 175, 203].

Moreover, massive amounts of blog entries and Twitter data can be analysed empirically [173]. Patterns of human behaviour are characterised from several perspectives. Periodic patterns and power law behaviours have been found. Yakovenko [221] proposes an agent-based model to explain conservative quantity of money and debt, and examines relationships between economic activity and energy consumption. He suggests that money and energy will be the key factors shaping the future of human civilisation.

Some researchers investigate transportation from a physical point of view [80, 105, 224]. According to the study by Guimerà and Amaral [81], the worldwide airport network has properties of a small-world network. The degree and betweenness centrality distributions exhibit the power-law decay. Jung et al. [105] examine the traffic flows of the Korean highway system. They show that the traffic flow between two cities forms a gravity model.

1.5.5 Data Engineering and Computer Sciences

Query log analysis of a web search engine reveals human typical behaviour [14, 100]. According to Broder [28], search queries reveal three types of user intents: (1) "navigational" (the user wants to reach a particular website), (2) "informational" (the user wants to find a piece of information on the web), and (3) "transactional" (the user wants to perform a web-mediated task). When do people send queries to a search engine, and what types of queries do they send? This tendency has a strong correlation with real society and economy.

Jansen et al. [100] studied 1,005,296 real-time search queries during the 190 days that originated from 43,140 unique IP addresses. They examined intradaily seasonality, weekly dependence, query length and term frequency. They found that real-time query logs can be a good representative quantity to characterise users, linking users' zip codes to U.S. census data. Weber and Jaimes found that the Yahoo! query logs provide a good demographic description of the U.S. population and that there are different segments in the topics that they search for as well as distinctions in their search behaviour [217]. Kato et al. [111] study how users of a web search engine use query suggestions. They analyse three kinds of data sets obtained from Microsoft's Bing search engine, comprising approximately 126 million unique queries, 876 mil-

lion query suggestions and 306 million action patterns of users. They show that query suggestions are often used (1) when the original query is a rare query; (2) when the original query is a single-term query; (3) when query suggestions are unambiguous; (4) when query suggestions are generalisations or error corrections of the original query; and (5) after the user has clicked on several URLs in the first search result page. Mitamura and Yoshida [142] propose a method for analysing social interests from DNS query logs. They investigate reactions of Japanese people on the eco-point system and the switchover to the digital TV broadcasts using their proposed method.

The launch of the E-Stat, which is a portal site of the Japanese government [162], provides us with new technological means for a data-based understanding of our country. In principle, everyone can understand the state of our country based on demographic data. Furthermore, real-time demographic data has also been available since the technologies to collect human activities via personal mobile phones were developed [72]. In the near future, we will be able to visualise real-time demographics, both comprehensively and circumstantially. Gao et al. [65] study real anomalous events using country-wide mobile phone data, finding that information flow during emergencies is dominated by repeated communications. They show that human communications are both temporally and spatially localised following the onset of emergencies, indicating that social propagation is a primary means to propagate situational awareness. They further demonstrate that the observed communication patterns cannot be explained by inherent reciprocity in social networks, and are universal across different demographics.

Recently, several car navigation companies have launched autonomous sensory navigation services. As a result, these companies can collect real-time car traffic data via each car navigation terminal. By collecting data from many cars, one can find roads and points where traffic jams are occurring. Without constructing new infrastructure to collect traffic states, real-time traffic data can be accumulated through the development of Integrated Transport Systems (ITS). Based on this data, comprehensive analyses of traffic flows can be conducted in order to address the problem of traffic jams [5]. These and other recent developments in traffic measurement technologies have been driving the theoretical development of traffic control and modelling [88].

Web-based commerce systems enable us to purchase everything from books to electronic equipment via websites. The details of consumers and goods can be stored on the database engine of each website. If we can access this data, then we can, in principle, capture real-time demand and supply of all items that are traded via websites.

Analysing massive amounts of data on items that are sold via web commerce systems is expected to open a window to new economic theory and service engineering [50, 125, 151]. Data on hotel booking opportunities [177], international flight booking opportunities [83, 174], intercity passenger railway [40] and price comparison sites [145] have also been studied. POS is an abbreviation for point-of-sales, and all department stores and supermarkets have introduced this kind of system in order to ring up purchases at cash registers. As a result, retail sales can be managed in real-time, and data-centric operations can be performed. On the basis of these massive

amounts of data, new marketing methods have been developed. The statistical properties of expenditure in a single shopping trip show a power law distribution [144]. Comprehensive analysis of retail sales is one of the most promising directions to be followed in order to bridge the gap between microeconomics and macroeconomics.

1.5.6 Computational Social Science

Lazer et al. address researches of computational social science from several perspectives [127]. They mention studies on video recording and analysis, examination of group interactions through electronic communication, examination of face-to-face group interactions over time, macro communication patterns, tracking movement and Internet as computational social science researches. Moreover, they emphasise that in a computational social science, properly managing privacy issues should be essential. Anonymisation technique is applied to solve the privacy issue. However, research revealed the potential for de-anonymisation, based on the statistical power for the sheer quantity of data collected from each individual in the database [12].

Kosinski et al. [122] show that easily accessible digital records of behaviour can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age and gender.

King [116] suggests that it is necessary to build up the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology and research ethics and policy in order to make progress possible in analysing, understanding and addressing major societal problems. Specifically, he mentions that data sharing regarding privacy protection plays an important role to progress studies using social science data. He proposes that we need to develop a common, open-source, collaborative infrastructure that makes data analysis and sharing easy under interoperation across scholarly fields. Moreover, social scientists can use additional help from the legal community. Standard rules and data-use agreements need to be developed.

Golder and Macy [71] identified individual-level diurnal and seasonal mood rhythms in cultures across the globe, using data from millions of public Twitter messages. They found that individuals awaken in a good mood that deteriorates as the day progresses—which is consistent with the effects of sleep and circadian rhythm—and that seasonal change in baseline positive affect varies with change in day-length.

Many observations of the dynamics of pedestrian crowds, including various self-organisation phenomena, have been successfully examined. Helbing et al. [88] studied video recordings of the crowd disaster in Mina/Makkah during the Hajj in 1426H on 12 January 2006. They found two subsequent, sudden transitions from laminar to stop-and-go and turbulent flows, which question many previous simulation mod-

els. Johansson et al. [102] examine social force pedestrian model by evolutionary adjustment mechanism to video tracking data.

1.6 Conclusion

The concept of complexity has been discussed from several points of view. Complexity is determined by the relative degrees of freedom of observing systems and in observed systems. I have briefly explained a framework of data-centric sciences. Applied data-centric social sciences can be established from data-centric studies based on large amounts of data on socioeconomic-technological systems. This data can be obtained from our society. However, to overcome the complexity of socioeconomic-technological data, we need to develop mathematical concepts, algorithms, rich computer systems and integrated databases. To construct synthesised data on socioeconomic-technological systems, spatial-temporal axes should be employed. Various kinds of data collected from different paths can be synthesised with time and locations. Rich synthesised databases representing aspects of our society may help us to obtain deeper insights into our own socioeconomic-technological systems.

I have also suggested that these studies are associated with applied data-centric social sciences. These applied data-centric social sciences include several research topics such as mathematical concepts, algorithms, computer systems, databases and modelling for data analysis.

Mathematical concepts can help us to develop automated algorithms to extract knowledge on socioeconomic-technological systems. These algorithm and computation systems help us to collect, store and analyse large amounts of data. It is also crucial to construct synthesised data sets from uncensored data obtained from different data sources. Furthermore, models can then become a guide for us to understand observed systems, connecting physical mechanisms with observations.

The property that these studies seem to have in common is their ability to overcome complexity in socioeconomic-technological systems by using massive amounts of data and vast computations. Copious amounts of data on human activities are collected by means of ICT, and vast amounts of computation for this data are conducted for the purposes of searching, matching, visualising and extracting.

This book is organised into four parts, including this introduction. In Part I, background and motivation of applied data-centric social sciences were shown. Part II shows mathematical methods in order to conduct data analysis of socioeconomic-technological systems. Fundamental methods of statistical inference, time series analysis, network analysis, and spatial analysis are explained. Part III contains several exemplar studies of socioeconomic-technological systems. Part IV addresses future work of applied data-centric social sciences.

The following list indicates potential fields of data-centric social sciences:

- Financial markets

- E-commerce
- Transportation
- Telecommunications
- Consumer products and Retail
- Demography
- Social network
- Search log
- Tourism
- Natural disaster prevention

I show several exemplar studies from some of these topics as examples of applied data-centric social sciences. The data is now available in various types of fields in our society, so, using additional concepts and methods, data from some specific fields is analysed. These exemplar studies should provide readers with much information to solve their own problems.

References

1. Acemoglu, D., Chernozhukov, V., Yildiz, M.: Fragility of asymptotic agreement under bayesian learning (2009). http://www.dklevine.com/archive/refs4814577000000000139.pdf
2. Albeverio, S., Jentsch, V., Kantz, H. (eds.): Extreme Events in Nature and Society. Springer, Berlin (2006)
3. Alfarano, S., Lux, T., Wagner, F.: Estimation of a simple agent-based model of financial markets: an application to Australian stock and foreign exchange data. Phys. A **370**, 38–42 (2006)
4. Antal, T., Redner, S., Sood, V.: Evolutionary dynamics on degree-heterogeneous graphs. Phys. Rev. Lett. **96**, 188104 (2006)
5. Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N.: Dynamic traffic demand prediction using conventional and emerging data sources. IEE Proc. Intell. Transp. Syst. **153**, 97–104 (2006)
6. Aoyama, H., Fujiwara, Y., Iyetomi, H., Sato, A.H. (eds.): Econophysics 2011— the Hichhiker's guide to the economy. In: Proceedings of the YITP Workshop on Econophysics. Oxford University Press, Oxford (2012). http://ptps.oxfordjournals.org/content/194.toc
7. Aoyama, H., Fujiwara, Y., Ikeda, Y., Iyetomi, H., Souma, W., Yoshikawa, H.: Econophysics and Companies: Statistical Life and Death in Complex Business Networks. Cambridge University Press, Cambridge (2010)
8. Arthur, W.B.: Increasing returns in the new world of business. Harvard Business Review (1996). http://hbr.org/1996/07/increasing-returns-and-the-new-world-of-business/ar/1. Accessed 3 March 2014
9. Aruka, Y.: Complexities of Production and Interacting Human Behaviour. Physica-Verlag, Heidelberg (2011)
10. Askitas, N., Zimmermann, K.F.: Google econometrics and unemployment forecasting. Appl. Econ. Q. **55**(2), 107–120 (2009)
11. Axelrod, R.M.: The dissemination of culture: a model with local convergence and global polarization. J. Conflict Resolut. **41**(2), 203–226 (1997)
12. Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. Commun. of the ACM **54**(12), 133–141 (2011)

13. Baek, Y., Ha, M., Jeong, H.: Absorbing states of zero-temperature glauber dynamics in random networks. Phys. Rev. E 85, 031123 (2012). http://link.aps.org/doi/10.1103/PhysRevE.85.031123
14. Baeza-Yates, R., Maarek, Y.: Usage Data in Web Search: Benefits and Limitations. In: Ailamaki, A., Bowers, S. (eds.) Scientific and Statistical Database Management: Lecture Notes in Computer Science, vol. 7338, pp. 495–506. Springer, Berlin (2012)
15. Bahr, D.B., Passerini, E.: Statistical mechanics of opinion formation and collective behavior: Micro-sociology. J. Math. Sociol. **23**, 1–27 (1998)
16. Bass, F.: A new product growth model for consumer durables. Manage. Sci. **15**, 215–227 (1969)
17. Battiston, S., Puliga, M., Kaushik, R., Tasca, P., Caldarelli, G.: DebtRank: too central to fail? financial networks, the FED and systemic risk. Sci. Rep. **2**, 541 (2012)
18. BBC: Population seven billion: UN sets out challenges. http://www.bbc.co.uk/news/world-15459643. Accessed 15 Aug 2013
19. Bettencourt, L.M.A., Lobo, J., Helbing, D., Kühnert, C., West, G.B.: Growth, innovation, scaling, and the pace of life in cities. Proc. Natl. Acad. Sci. USA **104**, 7301–7306 (2007)
20. Bettencourt, L.M.A., Lobo, J., Strumsky, D., West, G.B.: Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. PLoS ONE **5**(11), e13541 (2010)
21. Biswas, S., Sen, P.: Effect of the nature of randomness on quenching dynamics of the Ising model on complex networks. Phys. Rev. E **84**, 066107 (2011)
22. Blank, A., Solomon, S.: Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). Phys. A **287**, 279–288 (2000)
23. Bonanno, G., Caldarelli, G., Lillo, F., Micciché, S., Vandewalle, N., Mantegna, R.N.: Networks of equities in financial markets. Eur. Phys. J. B **38**, 363–371 (2004)
24. Bora, N., Zaytsev, V., Chang, Y.H., Maheswaran, R.: Spatiotemporal patterns in social networks. In: AAAI Technical Report FS-13-05, pp. 8–15 (2013)
25. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I.: Web search queries can predict stock market volumes. PLoS ONE **7**(7), e40014 (2012)
26. Bouchaud, J.P., Potters, M.: Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management. Cambridge Unversity Press, Cambridge (2000)
27. Bravo, G., Squazzoni, F., Boero, R.: Trust and partner selection in social networks: an experimentally grounded model. Soc. Netw. **34**, 481–492 (2012)
28. Broder, A.: A taxonomy of web search. SIGIR Forum **36**(2), 3–10 (2002)
29. Butler, D.: When google got flu wrong. Nature **494**(7436), 155–156 (2013)
30. Cambria, E., Grassi, M., Hussain, A., Havasi, C.: Sentic computing for social media marketing. Multimedia Tools Appl. **59**(2), 557–577 (2012)
31. Carrassi, A., Ghil, M., Trevisan, A., Uboldi, F.: Data assimilation as a nonlinear dynamical systems problem: stability and convergence of the prediction-assimilation system. Chaos **18**, 023112 (2008)
32. Castellani, B., Hafferty, F.W.: Sociology and Complexity Science A New Field of Inquiry. Springer, Berlin (2009)
33. Castellano, C., Loreto, V., Barrat, A., Cecconi, F., Parisi, D.: Comparison of voter and glauber ordering dynamics on networks. Phys. Rev. E **71**, 066107 (2005). http://link.aps.org/doi/10.1103/PhysRevE.71.066107
34. Castellano, C.: Social influence and the dynamics of opinions: the approach of statistical physics. Manag. Decis. Econ. **33**(5–6), 311–321 (2012). http://dx.doi.org/10.1002/mde.2555
35. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81**(2), 591–646 (2009)
36. Chakrabarti, B.K., Chakraborti, A., Chatterjee, A. (eds.): Econophysics and Sociophysics: Trends and Perspectives. Wiley, Belrin (2007)
37. Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States, http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf. Accessed 21 Oct 2013

38. Chapel, L., Castello, X., Bernard, C., Deffuant, G., Eguiluz, V.M., Martin, S., Miguel, M.S.: Viability and resilience of languages in competition. PLoS ONE **5**(1), e8681 (2010)
39. Chen, P., Redner, S.: Majority rule dynamics in finite dimensions. Phys. Rev. E **71**(3 Pt 2A), 036101 (2005)
40. Chi-Kang, L., Tzuoo-Ding, L., Chao-Hui, L.: Pattern analysis on the booking curve of an inter-city railway. J. East. Asia Soc. Transp. Stud. **6**, 303–317 (2005)
41. Clarke, A.C.: Profiles of the Future: An Inquiry into the Limits of the Possible. Harper & Row, New York (1962)
42. Czaplicka, A., Hołyst, J.A.: Modeling of internet influence on group emotion. Int. J. Modern Phys. C **23**(03), 1250020 (2012). http://www.worldscientific.com/doi/abs/10.1142/S0129183112500209
43. Da, Z., Engelberg, J., Gao, P.J.: In search of attention. J. Finance **66**(5), 1461–1499 (2011)
44. Daley, R.: Atmospheric Data Analysis. Cambridge University Press, Cambridge (1991)
45. Daley, R.: Atmospheric data assimilation. J. Meteorol. Soc. Japan **75**, 319–329 (1997)
46. D'Amuri, F., Marcucci, J.: "Google it!" Forecasting the US unemployment rate with a Google job search index. ISER Working Paper Series 2009–32 (2009)
47. Davenport, T.H., Patil, D.J.: Data scientist: the sexiest job of the 21st century. Harvard Bus. Rev. **90**, 70–76 (2012)
48. Davison, D.B., Chen, E.: A brief introduction to the internet. Comput. Geosci. **21**(6), 731–735 (1995)
49. Demographia: World urban areas & population projections. http://www.demographia.com/db-worldua.pdf. Accessed 15 Nov 2013
50. Deschâtres, F., Sornette, D.: Dynamics of book sales: endogenous versus exogenous shocks in complex networks. Phys. Rev. E **72**, 016112 (2005)
51. Dixit, A.K., Weibull, J.W.: Political polarization. Proc. Natl. Acad. Sci. USA **104**(18), 7351–7356 (2007)
52. Drożdż, S., Kwapień, J., Oświęcimka, P., Rak, R.: The foreign exchange market: return distributions, multifractality, anomalous multifractality and the Epps effect. New J. Phys. **12**, 105003 (2010)
53. Duan, W.Q., Stanley, H.E.: Cross-correlation and the predictability of financial return series. Phys. A **390**, 290–296 (2011)
54. Economist: Data, data everywhere (2010). http://www.economist.com/node/15557443. Accessed 12 July 2013
55. Eisler, Z., Kertész, J.: Scaling theory of temporal correlations and size-dependent fluctuations in the traded value of stocks. Phys. Rev. E **73**, 046109 (2006)
56. Embrechts, P., Resnick, S.I., Samorodnitsky, G.: Extreme value theory as a risk management tool. North Am. Actuarial J. **3**, 30–41 (1999)
57. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. Springer, Berlin (2000)
58. Epstein, J.M.: Why model? J. Artif. Soc. Soc. Simul. **11**(4), 12 (2008)
59. Evans, J.A., Foster, J.G.: Metaknowledge. Science **331**(6018), 721–725 (2011). doi:10.1126/science.1201765
60. Flightradar24. http://www.flightradar24.com. Accessed 31 Jan 2014
61. Fujie, R., Aihara, K., Masuda, N.: A model of competition among more than two languages. J. Stat. Phys. **151**(1–2), 289–303 (2013)
62. Gabaix, X.: Zipf's law for cities: an explanation. Q. J. Econ. **114**, 739–767 (1999)
63. Galam, S.: Minority opinion spreading in random geometry. Eur. Phys. J. B **25**(4), 403–406 (2002)
64. Gao, W., Yoshinaga, N., Kaji, N., Kitsuregawa, M.: Collective sentiment classification based on user leniency and product popularity. In: W. Gao, N. Yoshinaga, N. Kaji, M. Kitsuregawa (eds.) Collective Sentiment Classification based on User Leniency and Product Popularity, pp. 357–365. Department of English, National Chengchi University (2013). http://id.nii.ac.jp/0069/00024244

65. Gao, L., Song, C., Gao, Z., Barabási, A.L., Bagrow, J.P., Wang, D.: Quantifying information flow during emergencies. Sci. Rep. **4**, 3997 (2014)
66. Garas, A., Garcia, D., Skowron, M., Schweitzer, F.: Emotional persistence in online chatting communities. Sci. Rep. **2**, 402 (2012). http://dx.doi.org/10.1038/srep00402
67. Garas, A., Argyrakis, P.: A network approach for the scientific collaboration in the European framework programs. Europhys. Lett. **84**(6), 68005 (2008)
68. Garibaldi, U., Scalas, E.: Finitary Probabilistic Methods in Econophysics. Cambridge University Press, Cambridge (2010)
69. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1015 (2009)
70. Glauber, R.J.: Time dependent statistics of the Ising model. J. Math. Phys. **4**, 294–307 (1963)
71. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science **333**(6051), 1878–1881 (2011). doi:10.1126/science.1202775
72. González, M.C., Hidalgo, C.A., Barabási, A.L.: Understanding individual human mobility patterns. Nature **453**, 779–782 (2008)
73. González-Val, R., Lanaspa, L., Sanz-Gracia, F.: Gibrat's law for cities, growth regressions and sample size. Econ. Lett. **118**, 367–369 (2013)
74. Goodchild, M.F.: Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. Int. J. Spat. Data Infrastruct. Res. **2**, 24–32 (2007)
75. Goodman, N.: Fact, Fiction, and Forecast. Harvard University Press, Cambridge (1955)
76. Goodwin, R.: The nonlinear accelerator and the persistence of business cycles. Econometrica **19**, 1–17 (1951)
77. Google Flu Trend. http://www.google.org/flutrends. Accessed 5 Jan 2014
78. Granovetter, M.S.: The strength of weak ties. Am. J. Sociol. **78**, 1360–1380 (1973)
79. Granovetter, M.S.: Threshold models of collective behavior. Am. J. Sociol. **83**(6), 1420–1443 (1978)
80. Guimer, R., Mossa, S., Turtschi, A., Amaral, L.A.N.: The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc. Natl. Acad. Sci. USA **102**(22), 7794–7799 (2005). http://www.pnas.org/content/102/22/7794.abstract
81. Guimerà, R., Amaral, L.A.N.: Modeling the world-wide airport network. Eur. Phys. J. B **38**, 381–385 (2004)
82. Gworek, S., Kwapień, J., Drożdż, S.: Sign and amplitude representation of the forex networks. Acta Phys. Pol. A **117**, 681–687 (2010)
83. Haensel, A., Koole, G., Erdman, J.: Estimating unconstrained customer choice set demand: a case study on airline reservation data. J. Choice Model. **4**(3), 75–87 (2011). http://dx.doi.org/10.1016/S1755-5345(13),70043--5
84. Häggström, O.: Zero-temperature dynamics for the ferromagnetic Ising model on random graphs. Phys. A **310**(3–4), 275–284 (2002)
85. Handy, C.: 21 Ideas for Managers. Jossey-Bass, San Francisco (2000)
86. Hashimoto, Y., Ito, T.: Effects of Japanese macroeconomic statistic announcements on the dollar/yen exchange rate: high-resolution picture. J. Jpn. Int. Econ. **24**, 334–354 (2010)
87. Hegselmann, R., Flache, A.: Understanding complex social dynamics: a plea for cellular automata based modelling. J. Artif. Soc. Soc. Simul. **1**(3) (1998). http://jasss.soc.surrey.ac.uk/1/3/1.html
88. Helbing, D., Johansson, A., Al-Abideen, H.Z.: Dynamics of crowd disasters: an empirical study. Phys. Rev. E **75**, 046109 (2007)
89. Helbing, D.: Social Self-Organization. Springer, Belrin (2012)
90. Hickes, J.: A Contribution to the Theory of the Trade Cycle. Oxford University Press, Oxford (1950)
91. Himmler, F., Amberg, M.: Data integration framework for heterogeneous system landscapes within the digital factory domain. In: Procedia Engineering 69, 1138–1143 (2014). http://dx.doi.org/10.1016/j.proeng.2014.03.102
92. Holme, P., Newman, M.E.J.: Nonequilibrium phase transition in the coevolution of networks and opinions. Phys. Rev. E **74**(5 Pt 2), 056108 (2006)

93. Horvatic, D., Stanley, H.E., Podobnik, B.: Detrended cross-correlation analysis for non-stationary time series with periodic trends. Europhys. Lett. **94**, 18007 (2011)
94. Huang, X., Vodenska, I., Havlin, S., Stanley, H.E.: Cascading failures in bi-partite graphs: Model for systemic risk propagation. Scientific reports **3**, 1219 (2013)
95. Hui, D., Jackson, R.B.: Uncertainty in allometric exponent estimation: a case study in scaling metabolic rate with body mass. J. Theor. Biol. **249**(1), 168–177 (2007)
96. Hunt, R.: Internet-services, facilities, protocols and architecture. Comput. Commun. **20**, 1397–1411 (1998)
97. Inoue, J.I., Sazuka, N.: Queueing theoretical analysis of foreign currency exchange rates. Quant. Financ. **10**, 121–130 (2010)
98. Ishii, A., Arakaki, H., Matsuda, N., Umemura, S., Urushidani, T., Yamagata, N., Yoshida, N.: The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process. New J. Phys. **14**(6), 063018 (2012). www.stacks.iop.org/1367-2630/14/i=6/a=063018
99. Ising, E.: Beitrag zur theorie des ferromagnetismus. Zeitschrift für Physik **31**(1), 253–258 (1925). http://dx.doi.org/10.1007/BF02980577
100. Jansen, B.J., Liu, Z., Weaver, C., Campbell, G., Gregg, M.: Real time search on the web: Queries, topics, and economic value. Inf. Process. Manage. **47**(4), 491–506 (2011)
101. Jiang, S., Zhang, J., Ong, Y.S.: A multiagent evolutionary framework based on trust for multi-objective optimization. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems Volume 1, AAMAS '12, pp. 299–306. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2012). http://dl.acm.org/citation.cfm?id=2343576.2343619
102. Johansson, A., Helbing, D., Shukla, P.K.: Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. Adv. Complex Syst. **10**(2), 271–288 (2007)
103. Johnson, J.: Hypernetworks for policy design in systems of systems of systems. In: Glass, K., Colbaugh, R., Ormerod, P., Tsao, J. (eds.) Complex Sciences: Lecture Notes of the Institute for Computer Sciences. Social Informatics and Telecommunications Engineering, vol. 126, pp. 179–189. Springer, Cham (2013)
104. Jung, C.G.: Synchronicity: An Acausal Connecting Principle. Routledge, London (2006)
105. Jung, W.S., Wang, F., Stanley, H.E.: Gravity model in the korean highway. Europhys. Lett. **81**(4), 48005 (2008)
106. Kaitaniemi, P.: Testing the allometric scaling laws. J. Theor. Biol. **228**(2), 149–153 (2004)
107. Kaldor, N.: A model of the trade cycle. Econ. J. **50**, 78–92 (1940)
108. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82** (Series D), 35–45 (1960). www.cs.unc.edu/~welch/kalman/media/pdf/Kalman1960.pdf
109. Kaltwasser, P.R.: Uncertainty about fundamentals and herding behavior in the forex market. Phys. A **389**, 1215–1222 (2010)
110. Kasuga, Y., Ichikawa, M., Deguchi, H., Kanatani, Y.: A simulation model for analyzing the night-time emergency health care system in Japan. Dev. Bus.Simul. Experiential Learn. **38**, 171–181 (2011)
111. Kato, M.P., Sakai, T., Tanaka, K.: When do people use query suggestion? A query suggestion log analysis. Inf. Retrieval **16**(6), 725–746 (2013)
112. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: a review of the state-of-the-art. Inf. Fusion **14**(1), 28–44 (2013). http://dx.doi.org/10.1016/j.inffus.2011.08.001
113. Kholodilin, K.A., Podstawski, M., Siliverstovs, B.: Do google searches help in nowcasting private consumption? A real-time evidence for the US. KOF Working Papers 256 (2010)
114. Kietzmann, J., Canhoto, A.: Bittersweet! understanding and managing electronic word of mouth. J. Public Aff. **13**, 146–159 (2013)
115. Kim, H., Bearman, P.: The structure and dynamics of movement participation. Am. Sociol. Rev. **62**, 70–93 (1997)

116. King, G.: Ensuring the data-rich future of the social sciences. Science **331**(6018), 719–721. New York Science (2011). doi:10.1126/science.1197872
117. Kirman, A.: Ants, Rationality, and Recruitment. Q. J. Econ. **108**(1), 137–156 (1993). ideas.repec.org/a/tpr/qjecon/v108y1993i1p137-56.html
118. Kitagawa, G.: Data centric science for information society. In: Takayasu, H., Takayasu, M., Watanabe, T. (eds.) Econophysics Approaches to Large-Scale Business Data and Financial Crisis, pp. 211–225. Springer, Tokyo (2010)
119. Knight, F.H.: Risk, Uncertainty, and Profit. Houghton Mifflin, New York (1921)
120. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2008)
121. Korth, H.F.: Database research faces the information explosion. Commun. ACM **40**, 139–142 (1997)
122. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. USA **110**(15), 5802–5805 (2013)
123. Kuninaka, H., Matsushita, M.: Modelling of population migration to reproduce rank-size distribution of cities in Japan. In: Zhou, J. (ed.) Complex Sciences: Lecture Notes of the Institute for Computer Sciences. Social Informatics and Telecommunications Engineering, vol. 4, pp. 441–445. Springer, Berlin (2009)
124. Kydland, F.E., Prescott, E.C.: Time to build and aggregate flucuations. Econometrica **50**, 1345–1370 (1982)
125. Lambiotte, R., Ausloos, M.: Endo-vs. exo-genous shocks and relaxation rates in book and music sales. Phys. A **362**, 485–494 (2005)
126. Landauer, T.K.: How much do people remember? Some estimates of the quantity of learned information in long-term memory. Cognitive sci. **10**, 477–493 (1986)
127. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Alstyne, M.V.: Life in the network: the coming age of computational social science. Science **323**(5915), 721–723 (2009)
128. Lebaron, B.: Agent-Based Computational Finance. In: Schmedders, K., Judd, K.L. (eds.) Handbook of Computational Economics, pp. 1187–1233. Elsevier, Amsterdam (2006)
129. Lee, E.A.: Cps foundations. In: Proc. of the 47th Design Automation Conference (DAC), pp. 737–742. ACM (2010)
130. Leyvraz, F., Redner, S.: Scaling theory for migration-driven aggregate growth. Phys. Rev. Lett. **88**, 068301 (2002)
131. Li, J., Liu, F.: A proposed framework of ewom and etrust in online hotel booking: the influence of an e-intermediary. In: International Conference on Management and Service Science (MASS) 2011, pp. 1–4 (2011)
132. Liu, L.Z., Qian, X.Y., Lu, H.Y.: Cross-sample entropy of foreign exchange time series. Phys. A **389**, 4785–4792 (2010)
133. Mantegna, R.N.: Hierarchical structure in financial markets. Eur. Phys. J. B **11**, 193–197 (1999)
134. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2000)
135. Marsden, P.V., Campbell, K.E.: Measuring tie strength. Soc. Forces **63**, 482–501 (1984)
136. Marwell, G., Oliver, P.: The Critical Mass in Collective Action: A Micro-Social Theory. Cambridge University Press, New York (1993)
137. Masuda, N., Ohtsuki, H.: Evolutionary dynamics and fixation probabilities in directed networks. New J. Phys. **11**, 033012 (2009)
138. Meadows, D.H., Meadows, D.L., Randers, J., Behrens, W.W.I.: The Limits to Growth. Universe Books, New York (1972)
139. Micciché, S., Bonanno, G., Lillo, F., Mantegna, R.N.: Volatility in financial markets: stochastic models and empirical results. Phys. A **314**, 756–761 (2002)
140. Mira, J., Seoane, L.F., Nieto, J.J.: The importance of interlinguistic similarity and stable bilingualism when two languages compete. New J. Phys. **13**(3), 033007 (2011). http://groups.lis.illinois.edu/amag/langev/paper/mira2011importancegsc.html

141. Mira, J., Paredes, A.: Interlinguistic similarity and language death dynamics. Europhys. Lett. **69**(6), 1031–1034 (2005)
142. Mitamura, T., Yoshida, K.: Viewers' side analysis of social interests. In: Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, ICDMW '12, pp. 301–308. IEEE Computer Society, Washington, DC, USA (2012). http://dx.doi.org/10.1109/ICDMW.2012.28
143. Mizuno, T., Ishikawa, A., Fujimoto, S., Watanabe, T.: Power laws in firm productivity. In: Aoyama, H., Fujiwara, Y., Iyetomi, H., Sato, A.H. (eds.) Econophysics 2011—the Hichhiker's guide to the economy. Proceedings of the YITP Workshop on Econophysics. vol. 194, pp. 122–134. Oxford University Press, Oxford (2012). http://ptps.oxfordjournals.org/content/194/122.full.pdf
144. Mizuno, T., Toriyama, M., Terano, T., Takayasu, M.: Pareto law of the expenditure of a person in convenience stores. Phys. A **387**, 3931–3935 (2008)
145. Mizuno, T., Watanabe, T.: A statistical analysis of product prices in online market. Eur. Phys. J. B **76**, 501–505 (2010)
146. Mostafa, M.M.: An emotional polarity analysis of consumers airline service tweets. Soc. Netw. Anal. Mining **3**(3), 635–649 (2013). http://dx.doi.org/10.1007/s13278-013-0111-2
147. Mountain, D., Raper, J.F.: Modelling human spatio-temporal behaviour: a challenge for location-based services. In: Proc. of the 6th Internat. Conference on GeoComputation, University of Queensland, Brisbane, Australia, pp. 24–26 September (2001)
148. Nakano, S., Higuchi, T.: Estimation of a long-term variation of a magnetic-storm index using the merging particle filter. IEICE TRANSACTIONS on Information and Systems **E92-D**(7), 1382–1387 (2009)
149. Ni, X.H., Jiang, Z.Q., Gu, G.F., Ren, F., Chen, W., Zhou, W.X.: Scaling and memory in the non-poisson process of limit order cancelation. Phys. A Statistical Mechanics and its Applications **389**(14), 2751–2761 (2010). http://dx.doi.org/10.1016/j.physa.2010.02.040
150. Nishi, R., Masuda, N.: Collective opinion formation model under bayesian updating and confirmation bias. Phys. Rev. E **87**(6), 062123 (2013)
151. O'Connor, P., Höpken, W., Gretzel, U. (eds.): Information and Communication Technologies in Tourism 2008. Springer, Vienna (2008)
152. Ohnishi, T., Mizuno, T., Aihara, K., Takayasu, M., Takayasu, H.: Statistical properties of the moving average price in dollar-yen exchange rates. Phys. A **344**, 207–210 (2004)
153. Oliver, P.E.: Formal models of collective action. Ann. Rev. Sociol. **19**, 271–300 (1993)
154. Olson, M.J.: The Logic of Collective Action: Public Goods and the Theory of Goods. Harvard University Press, Cambridge (1965)
155. Pan, B., Li, X.: The long tail of destination image and online marketing. Ann. Tourism Res. **38**(1), 132–152 (2011)
156. Patriarca, M., Heinsalu, E.: Influence of geography on language competition. Phys. A **388**(2–3), 174–186 (2009)
157. Paul, T., Kimball, J.W., Zawodniok, M.J., Roth, T.P., McMillin, B.M.: Invariants as a unified knowledge model for cyber-physical systems. In: K.J. Lin, C. Huemer, M.B. Blake, B. Benatallah (eds.) Proceedings of IEEE International Conference on Service Oriented Computing and Applications (SOCA), pp. 1–8. IEEE (2011)
158. Paul, T., Kimball, J.W., Zawodniok, M.J., Roth, T.P., McMillin, B.M., Chellappan, S.: Unified invariants for cyber-physical switched system stability. Smart Grid, IEEE Trans **5**(1), 112–120 (2014)
159. Podobnik, B., Horvatic, D., Petersen, A.M., Stanley, H.E.: Cross-correlations between volume change and price change. Proc. Natl. Acad. Sci. USA **106**, 22079–22084 (2009)
160. Podobnik, B., Horvatic, D., Petersen, A.M., Njavro, M., Stanley, H.E.: Common scaling behavior in finance and macroeconomics. Eur. Phys. J. B **76**, 487–490 (2010)
161. Pontieri, L., Ursino, D., Zumpano, E.: An approach for the extensional integration of data sources with heterogeneous representation formats. Data Knowledge Engineering **45**(3), 291–331 (2003). http://www.sciencedirect.com/science/article/pii/S0169023X02001921

162. Portal Site of Official Statistics of Japan by National Statistics Center. http://www.e-stat.go.jp. Accessed 3 March 2014
163. Preis, T., Golke, S., Paul, W., Schneider, J.J.: Multi-agent-based order book model of financial markets. Europhys. Lett. **75**(3), 510–516 (2006). http://stacks.iop.org/0295-5075/75/i=3/a=510
164. Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using google trends. Sci. Rep. **3**, 1684 (2013)
165. Railsback, S.F., Lytinen, S.L., Jackson, S.K.: Agent-based simulation platforms: Review and development recommendations. Simulation **82**(9), 609–623 (2006). http://sim.sagepub.com/content/82/9/609.abstract
166. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: explorations in urban data collection. Pervasive Comput. **6**, 30–38 (2007)
167. Rebitzky, R.R.: The influence of fundamentals on exchange rates: findings from analyses of news effects. J. Econ. Surv. **24**, 680–704 (2010)
168. Richards, L.: Handling Qualitative Data. Sage publications, London (2005)
169. Risken, H.: The Fokker-Planck Equation: Methods of Solution and Applications. Springer, Berlin (1989)
170. Rogers, R.: Digital methods. MIT Press, Cambridge (2010)
171. Sabherwal, R., Becerra-Fernandez, I.: Business Intelligence. Wiley, Hoboken (2011)
172. Samuelson, P.A.: Interactions between the multiplier analysis and the principle of acceleration. Rev. Econ. Stat. **21**, 75–78 (1939)
173. Sano, Y., Yamada, K., Watanabe, H., Takayasu, H., Takayasu, M.: Empirical analysis of collective human behavior for extraordinary events in the blogosphere. Phys. Rev. E **87**, 012805 (2013). http://link.aps.org/doi/10.1103/PhysRevE.87.012805
174. Sato, A.H.: Japanese international air travel: the relationship between flight ticket price and geodesic distance. In: Proceedings 2012 IEEE World Congress on Computational Intelligence (Brisbane), pp. 2821–2826 (2012)
175. Sato, A.H., Takayasu, H.: Dynamic numerical models of stock market price: from microscopic determinism to macroscopic randomness. Phys. A **250**, 231–252 (1998)
176. Sato, A.H., Hołyst, J.A.: Characteristic periodicities of collective behavior at the foreign exchange market. Eur. Phys. J. B **62**, 373–380 (2008)
177. Sato, A.H.: Patterns of regional travel behavior: an analysis of Japanese hotel reservation data. Int. Rev. of Finan. Anal. **23**, 55–65 (2012)
178. Sazuka, N., Inoue, J., Scalas, E.: The distribution of first-passage times and durations in forex and future markets. Phys. A **388**, 2839–2853 (2009)
179. Schelling, T.C.: Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities. J. Conflict Resolut. **17**, 381–428 (1973)
180. Schumpeter, J.A.: The Theory of Economic Development. Transaction Publishers, New Brunswick (1983)
181. Schweitzer, F.: Brownian Agents and Active Particles. Springer, Berlin (2003)
182. Shiozawa, Y., Matsui, H., Taniguchi, K., Nakajima, Y., Koyama, Y., Hashimoto, F.: Artificial Market Experiments with the U-Mart System. Springer, Tokyo (2008)
183. Shou, L., Wu, S.: Supporting efficient social media search in cyber-physical web. IEEE Data Eng. Bull. **36**(3), 83–90 (2013)
184. Sinha, S., Chatterjee, A., Chakraborti, A., Chakrabarti, B.K.: Econophysics: An Introduction. Wiley, Berlin (2010)
185. Smith, G.P.: Google internet search activity and volatility prediction in the market for foreign currency. Finance Res. Lett. **9**(2), 103–110 (2012)
186. Soni, S., Mehta, S., Hans, S.: Towards providing data validation as a service. In: Services Computing (SCC), 2012 IEEE Ninth International Conference on, pp. 570–577 (2012). doi:10.1109/SCC.2012.82
187. Sood, V., Redner, S.: Voter model on heterogeneous graphs. Phys. Rev. Lett. **94**(17), 178701 (2005)

188. Sornette, D.: Why Stock Markets Crash: Critical Events in Complex Financial Systems. Princeton University Press, Princeton (2003)
189. Souza, S.R., Goncaleves, S.: Dynamical model for competing opinions. Phys. Rev. E **85**, 056103 (2012)
190. Spirin, V., Krapivsky, P.L., Redner, S.: Fate of zero-temperature Ising ferromagnets. Phys. Rev. E **63**, 036118 (2001). http://link.aps.org/doi/10.1103/PhysRevE.63.036118
191. Swearingen, C.D., Ripberger, J.T.: Google Insights and U.S. senate elections: does search traffic provide a valid measure of public attention to political candidates? Social Science Quarterly (2014). (In press)
192. Takahashi, S., Sallach, D., Rouchier, J. (eds.): Advancing Social Simulation: The First World Congress. Springer, Tokyo (2007)
193. Takayasu, H. (ed.): The Advent of Econophysics. Spriger, Tokyo (2002)
194. Takayasu, H. (ed.): Practical Fruits of Econophysics. Springer, Tokyo (2006)
195. Taleb, N.N.: The Black Swan: The impact of the highly improbable. Random House, New York (2007)
196. Tan, K.C., Lee, T.H., Khor, E.F.: Evolutionary algorithms for multi-objective optimization: performance assessments and comparisons. In: Proceedings of the 2001 Congress on Evolutionary Computation 2001, vol. 2, pp. 979–986 (2001). doi:10.1109/CEC.2001.934296
197. Tanuma, H., Deguchi, H., Shimizu, T.: Soars: Spot oriented agent role simulator design and implementation. In: Terano, T., Kita, H., Kaneda, T., Arai, K., Deguchi, H. (eds.) Agent-Based Simulation: From Modeling Methodologies to Real-World Applications, Agent-Based Social Systems, vol. 1, pp. 1–15. Springer, Tokyo (2005)
198. Terano, T.: The shape of experiment-based management science to come (2012). The 1st General conference on emerging arts of research on management and administration (GEAR), http://www.trn.dis.titech.ac.jp/GEAR/pdf/Terano2012GEAR.pdf
199. Tesfatsion, L.: Agent-based computational economics: growing economies from the bottom up. Artif. Life **8**, 55–82 (2002)
200. Tesfatsion, L., Judd, K.L. (eds.): Handbook of Computational Economics. Elsevier, Amsterdam (2006)
201. Tomasello, M.V., Napoletano, M., Garas, A., Schweitzer, F.: The rise and fall of R&D networks (2013). http://arxiv.org/abs/1304.3623
202. Tsubaki, H.: Valuation of partly disclosed datasets for prediction. 2013 IEEE 13th International Conference on Data Mining Workshops pp. 733–734 (2013). http://doi.ieeecomputersociety.org/10.1109/ICDMW.2013.148
203. Tuncay, C.: Socioeconophysics: Opinion dynamics for number of transactions and price, a trader based model. Int. J. Mod. Phys. **17**, 1495–1500 (2006)
204. UK data archive. http://www.data-archive.ac.uk. Accessed 3 Jan 2014
205. Vazquez, F., Eguíluz, V.M., Miguel, M.S.: Generic absorbing transition in coevolution dynamics. Phys. Rev. Lett. **100**(10), 108702 (2008)
206. Vazquez, F., Castello, X.: Agent based models of language competition: macroscopic descriptions and order-disorder transitions. J. Stat. Mech. Theory Exp. **2010**(04), P04007 (2010)
207. Verdone, R. (ed.): Wireless Sensor Networks. Springer, Berlin (2008)
208. Vlastakis, N., Markellos, R.N.: Information demand and stock market volatility. J. Banking Finance **36**(6), 1808–1821 (2012)
209. Vosen, S., Schmidt, T.: Forecasting private consumption: survey-based indicators vs. google trends. J. Forecast. **30**(6), 565–578 (2011)
210. Von Bertalanffy, L.: General System Theory: Foundations, Development Applications. George Braziller, New York (1969)
211. Von Foerster, H.: Observing Systems. Intersystems Publications, California (1984)
212. Von Neumann, J.: The Computer and the Brain. Yale Univesity Press, New Haven (1958)
213. Wang, F., Shieh, S.J., Havlin, S., Stanley, H.E.: Statistical analysis of the overnight and daytime return. Phys. Rev. E **79**, 056109 (2009)
214. Watanabe, H., Takayasu, H., Takayasu, M.: Relations between allometric scalings and fluctuations in complex systems: the case of Japanese firms. Phys. A **392**, 741–756 (2013)

215. Watts, D.J.: A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. USA **99**(9), 5766–5771 (2002)
216. Watts, D.J., Dodds, P.S.: Influentials, networks, and public opinion formation. J. Consu. Res. **34**, 441–458 (2007)
217. Weber, I., Jaimes, A.: Who uses web search for what? And how? In: WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining, pp. 15–24. ACM (2011)
218. Wen, H.J., Chen, H.G., Hwang, H.G.: E-commerce web site design: strategies and models. Inf. Manage. Comput. Sec. **9**, 5–12 (2001)
219. Wymbs, C.: How e-commerce is transforming and internationalizing service industries. J. Serv. Mark. **14**, 463–477 (2000)
220. Xiang, Z., Pan, B.: Travel queries on cities in the united states: implications for search engine marketing for tourist destinations. Tourism Manage. **32**(1), 88–97 (2011)
221. Yakovenko, V.M.: Statistical mechanics of money, income, debt, and energy consumption. Sci. Culture **76**(9–10), 430–436 (2010)
222. Yamauchi, T., Hayashi, Y., Nakano, Y.I.: Searching emotional scenes in TV programs based on twitter emotion analysis. In: Ozok, A.A., Zaphiris, P. (eds.) Online Communities and Social Computing. Lecture Notes in Computer Science, vol. 8029, pp. 432–441. Springer, Berlin Heidelberg (2013)
223. Yang, C., Kurahashi, S., Ono, I., Terano, T.: Pattern-oriented inverse simulation for analyzing social problems: Family strategies in civil service examination in imperial china. Adv. Complex Syst. **15**, 1250038 (2012)
224. Zanin, M., Lacasa, L., Cea, M.: Dynamics in scheduled networks. Chaos **19**(2), 023111 (2009). http://scitation.aip.org/content/aip/journal/chaos/19/2/10.1063/1.3129785
225. Zhang, J., Yu, T.: Allometric scaling of countries. Phys. A **389**, 4887–4896 (2010)
226. Zhang, M., Gong, T.: Principles of parametric estimation in modeling language competition. Proc. Natl. Acad. Sci. USA **110**(24), 9698–9703 (2013)
227. Zimper, A., Ludwig, A.: On attitude polarization under bayesian learning with non-additive beliefs. J. Risk Uncertainty **39**(2), 181–212 (2009)

Chapter 2
Framework

Abstract A framework of the applied data-centric social sciences is based on data-centric science. A methodology of data-centric science is very common and applicable to all the types of sciences. In this chapter, we will see a methodology used in applied data-centric sciences commonly.

2.1 Pipelines of Data-Centric Science

Generally, the data-centric investigation or data-driven study is constructed from the following steps:

- problem definition
- project design
- an explanatory data analysis
- data acquisition
- data collection
- data analysis
- interpretation
- decision-making

These steps construct a cycle to improve data quality, interpretation adequateness and effectiveness of decision-making. In order to understand the data-generating mechanism, we also should visit actual spots where the data are generated and confirm correspondence between the data and objects or concepts which they express. In general, the problem and project are unknown firstly. We may not clearly understand the problem which we need to solve and the project where we should work. To understand them, it is useful to come in touch with data of the problem or of the field where the project will be built. This type of activity is called *explanatory data analysis* [34].

In both the inductive and deductive approaches, in general, we face the so-called chicken-and-egg problem. This is a kind of causality dilemma. The problem definition

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_2, © Springer Japan 2014

and the project building sometimes face the causality dilemma. This means that we need to build a project to understand details of the problem related to the project. The abductive approach may solve this causality dilemma. The abductive approach is defined as an approach to start to think the problem on a hypothesis. A plenty of data often becomes a starting point of the hypothesis. The explanatory data analysis provides us with knowledge on phenomena and characteristics of the problem.

2.2 Purpose, Goal and Proposal

We need a purpose or purposes for our research activity in order to justify our own activity. When we start our research or project, we have to ask ourselves whether we can improve our society, add new information to existing studies, solve a societal problem or propose a new policy to decision-makers. Our activity may simply create knowledge on a societal issue. Then, we need to consider how to contribute to our society by increasing knowledge on the societal issue.

The goal is different from the purpose. The goal of our research activity or our project should be concrete with some quantitative measures. For example, in the case of a business, the goal should be defined as a measurable improvement such as the number of consumers, the duration time for production, and so on. In the case of the academic research, the goal should find new things and/or propose new concepts or methods with some quantitative manners. How many or how much do we improve the activity or clarify the phenomena? To do so, we firstly need to grasp our current situation and determine an area where we can make a difference.

2.3 Project Design

In order to design our project or research activity, we must ask ourselves the following questions again and again during the project:

- What is our question?
- How do we ask and solve the question?
- What data do we need to answer the question?

We often start our project without any concrete goals. However, such a launch seems to create some problems during our own research activity. For example, imagine that we do not know what we should achieve and how to examine the issue. How do we feel about this situation?

Actually, we need to find a goal for our activity. Asking ourselves several questions may lead us to a concrete goal. To find an adequate question, a bird's eye view of the problem or phenomenon may help us. For example, we can ask ourselves as follows:

1. What is our research field?
2. Do we find any gaps between existing studies and general questions?
3. What is our focus?

4. What kinds of questions can we ask in our focus?
5. What can we expect to contribute based on our resources and skills?
6. Can we find any relationship between our standing point and the questions?

These questions may also help us to find a concrete goal for our project:

- What types of data do we need?
- How many or much data do we need?
- What types of data do we need to reach our goal?

Through these questions, we find a way to investigate our object to reach our goal.

How do you feel from these questions? You may not find any concrete answers to these questions. The main reasons why you cannot find any answers are because:

- a lack of information in the fields
- a lack of knowledge on the problem
- a lack of skills to solve the problem
- a lack of resources of the research

Then, you need to have an experience to treat the (even small) data on the problem or the field at least to find a concrete answer. You can start your explanatory data analysis from acquiring a small amount of data related to the field which you want to contribute to. And then, you will be able to find better answers to the problem.

Furthermore, during our research activity, it is important for us to often check whether our activity is adequate? To do so, it is useful to record logs of our own activities. In fact, documents or memos of our own activities help us to confirm our research activity. The research diary may be useful for this purpose. We can write our activity in research notes with dates. The software and procedures for computations should be also recorded. We need to check our activity during our research project repeatedly.

2.4 Data Acquisition

The data is recorded from some data-generating source. The data of society are currently available from web pages. Both personal and official web pages are a preliminary data source of our society. Electronic commerce systems are also sources of data for products and services. We can accumulate data on prices for goods and services from application programming interface (API) of some data providers. Data of financial markets, job opportunities, hotels, flights, traffic and so on are accumulated via Web API nowadays.

Web API is an application programming interface which can be used via the Internet. In Web API, there are several technologies to exchange commands and data between an API provider and users. Functions of natural language processing, geographical information systems (GIS), search engines and databases of e-commerce services are available as Web APIs. This list shows several examples:

- Yahoo! JAPAN text analytics WebAPI [37]
- Jalan vacant room information retrieval WebAPI [18]

- AB-ROAD travel retrieval WebAPI [2]
- Rakuten Web Service WebAPI [25]
- Google Translate WebAPI [14]

The secondary source of data is a sensor network. Several types of sensors have recently become available. Some of them can send data to a database server via the Internet directly. We require sensors that convert physical parameters to electrical signals. The sensor signals are converted into a form that can be converted to digital values. Analog-to-digital converters are included in the sensors. The sensors are connected with one another through wired or wireless network. This is sometimes referred to as Internet of things (IoT).

Machine-to-machine (M2M) solution is one of the implementations of IoT, which is provided from several vendors. Functional requirements of the M2M application are as follows:

- There are data that can be exchanged between a device and a server.
- There are device management capabilities provided by an M2M application.
- There are different components of which an M2M application is made.

Data management of the M2M application includes hierarchical structure of data elements. The data type can be associated with the data elements. Primitive data types such as string, integer, double, date, Boolean and byte array are supported. Users can define constraints for the data, identify the protocol to be used when exchanging a given data element and configure parameters to protocols. There are some commands that can be sent by a server to a device, and sets of events that can be sent by a device to a server.

2.5 Data Collection

Data collection is the process of gathering information. In the data collection, several types of sampling methods are known:

- simple random sampling
- systematic sampling
- snowball sampling
- comprehensive sampling

A simple random sampling means that we obtain a subset of individuals chosen from a larger set. Each piece of data is chosen randomly and has the same probability to be chosen at any stages.

A systematic sampling is to sample data according to some ordering scheme and then select elements at regular intervals through that ordered list, for example, selecting every 10th name from the telephone directory.

A snowball sampling is often used in sociology and statistics research. Snowball sampling is non-probability sampling where existing data recruits the potential data

which will be sampled in the future. Therefore, the sample group appears to grow like a rolling snowball. For example, suppose that we accumulate data of web pages from the World Wide Web. In this case, firstly, we choose a web page. Next, we select a page from a link included in the sampled web page. Repeating this procedure, we eventually collect data of web pages.

A comprehensive sampling means that we obtain all the data that we can cover. If we have sufficient computer resources and time, then we can conduct the comprehensive sampling.

The data are stored in computer systems as digital files such as CSV, TSV, XML and so on. These files can be inserted into database servers, which play an important role in data collection. A database management system (DBMS) is at the core of data collection. Some types of DBMS are recently available:

- relational database
- XML database
- object-oriented database
- document-oriented database

We need to handle several types of databases at the same time. In the data collection, we may need to determine the area of the data. A relational database management system (RDBMS) have a high affinity with CSV and TSV formats. XML formats can be transformed to CSV or TSV formats and can be handled by RDBMS. XML databases can be used to handle XML-formatted data directly.

Furthermore, we need to carefully consider the way to prevent data loss. The data loss badly affects results of data analysis and generate additional data acquisition and costs. Intentional and accidental deletion of files or data damages collections of data. Using a journaling file system and Redundant Arrays of Independent Disks (RAID) storage can protect against some types of software and hardware failure. Regular data backups are an effective method to recover the data from data-loss events. In fact, user errors or system failures cannot be prevented by regular backups, but we may quickly recover the system from such failures if we keep several versions of backups.

2.6 Data Validation

Data validation is one of the most important but the most time-consuming tasks [28]. Without clean data, data analysis and optimisation tools cannot work well. Data analysis and optimisation solutions always assume the presence of correct data. In many cases presence/inference of wrong data is even worse than absence of the data, and a harmful effect in decision-making will happen. Therefore, it is an important step for any researcher to verify and validate the accuracy and adequateness of the data. There are several types of validation methods:

- multiplexing data sources
- consistency check
  - item count validation test
  - range validation test
- finding outliers

Multiplexing data sources may help us to find data inconsistency. For example, suppose that we use macroeconomic statistics such as population or GDP. Then, we should collect the same data from two institutions. We can compare the same data elements obtained from the difference institutions at least. If quantities in the elements are different from each other, then we can understand that one of them is wrong or contains some error. This technique is, of course, applicable to other areas than macroeconomic data.

Consistency check is a common method for several types of data. In this case, a physical model of data-generating mechanism is useful. For example, causality, time and space can be used for this purpose. There are two types of data errors: systematics errors and random errors. Systematics errors can result from bugs of software to generate data or procedures. Thus, when they occur at all, they occur repeatedly. Systematic errors can produce three types of errors: (1) too many data elements, (2) too few data elements and (3) classification of data elements. The primary action of the data validation is to identify the occasions when systematic errors happen. (1) and (2) can be checked if we count the number of data elements. This is called *item count validation test*. (3) can be confirmed by checking the types of data and range of data. This is called *range validation test*. The range validation test is done by checking that all records are within specified ranges.

Random errors are generated as input errors or judgement errors. In general, random errors occur intermittently. This type of error can be detected as an outlier from other values. Both the range validation and item count validation tests can be used to detect random errors.

Range validation test is sometimes useful if the data are numeric or one of several options. If a data element is out of range, then we can determine that it is wrong data. When we use geographical information, we can use distance from a position as a norm of data. We may find incorrect data as some outliers from a relation of feature to the distance. When we use time series data, time order can be used to check the data consistency. If the time order is contrary or missing, then we may find incorrect data or missing data.

Outliers are defined as a data point that extremely differs from other data points. Ben-Gal [5] classifies outlier detection methods into univariate statistical methods and multivariate outlier detection. The earliest univariate methods for outlier detection use the assumption of an underlying known distribution of the data. An outlier can be detected by using mean of values included in a dataset and their standard deviation. If we assume that the values are sampled from a normal distribution, then the probability where the samples appear between the mean minus three times the standard deviation and the mean plus three times the standard deviation is 99.9 %. Therefore, the values deviating from this range are detected as outliers. Barnett and Lewis [3]

showed statistical methods to identify outliers (Chauvenet's criterion, Grubbs's test for outliers [15], Peirce's criterion [23], Dixon's Q test [9], Thompson test). In the multivariate case, *Mahalanobis* distance [21] can be used. The *Mahalanobis* distance [21] for each multivariate data point $\boldsymbol{x}_s$ $(s = 1, \ldots, T)$ is defined as

$$M_s = \left((\boldsymbol{x}_s - \bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{V}_n^{-1} (\boldsymbol{x}_s - \bar{\boldsymbol{x}})\right)^{1/2}, \tag{2.1}$$

where $\boldsymbol{V}$ represents the sample covariance matrix defined as

$$\boldsymbol{V} = \frac{1}{T-1} \sum_{s=1}^{T} (\boldsymbol{x}_s - \bar{\boldsymbol{x}})(\boldsymbol{x}_s - \bar{\boldsymbol{x}})^{\mathrm{T}}, \tag{2.2}$$

and $\bar{\boldsymbol{x}}$ the sample mean. A large value of M_s for the s-th data point indicates that it is an outlier.

During the data analysis, we may often find some outliers. In this case, we should check whether the outliers are consistent with the mechanism to generate the data or not.

Data quality problems are recognised as important tasks in data engineering. Detecting and removing errors and inconsistencies from data improve the quality of data. These tasks are called "data cleaning". There is a big range of data cleaning commercial tools available in the market. Some of those are more generic in operation and others are solving a specific problem in a particular domain. Rahm and Do [24] also propose five phases of data cleaning approaches to construct an automated data-cleaning system for data warehousing:

- Data analysis
- Definition of transformation work flow and mapping rules
- Verification
- Transformation
- Backflow of cleaned data

The data analysis is needed in order to detect which kinds of errors and inconsistencies are to be removed. The definition of schema-related data transformations and mapping rules for data elements should be considered. The correctness and effectiveness of a transformation work flow and the transformation definitions should be tested and evaluated. The transformation steps are executed. After errors are removed, the cleaned data should also replace the dirty data in the original sources. To data quality management (DQM) to the data loaded in the system, we need define DQM rules that perform a variety of repair, clean up, and standardisation functions on incoming identity data values. These functions are implemented in recent Big Data

Analytics solutions such as IBM Netezza,[1] SAP Data Quality Management software[2] and Talend Enterprise Data Quality solution.[3]

2.7 Explanatory Data Analysis

One of the most important steps in the investigation is an explanatory analysis [34]. This is a kind of feasible study. The explanatory data analysis consists of the following steps:

1. visualise data and compute fundamental statistics
2. construct a model from ideas obtained from statistical analysis
3. estimate model parameters
4. validate or check an adequacy of the model with parameter estimates
5. interpret data using the estimated model
6. repeat 1–5 until we are satisfied with the interpretation

Concretely, this procedure can be drawn as

1. Make scatter plots between variables, draw time series, networks and spatial plots, and make a histogram from observations. From the plots, we can find some patterns and detect outliers of data. If we need a new axis of data, we define it or additionally start to collect data from environment. It is useful to compute descriptive statistics (mean, variance, quartile, skewness, and kurtosis). Changing granularity of data or spatio-temporal scales we need to compute these fundamental properties of data.
2. Applying methods of multivariate analysis (regression analysis, principal component analysis, spatial regression, and factor analysis) and time series analysis (autoregressive analysis) we need to determine relationship among variables and their temporal transitions (transition probabilities). These processes provide us with ideas on data generating mechanisms and stochastic models as an approximation of actual mechanism.
3. Repeating step 1 and step 3, we increase kinds of data and accumulate the number of observations as well as ideas of models for variables.
4. Realising the model, we attempt to estimate model parameters from observations. If we use a regression model, we will check goodness-of-fit of data for the model in terms of an explained variable and explanatory variables. A degree of freedom of the model is determined by using some criteria such as information criteria. Data bias (sampling bias, processing bias, and so on) should be taken into account in this step. During this step, we sometimes recognise a fault of data acquisition or data collection.

[1] IBM Netezza: http://www-01.ibm.com/software/data/netezza/.

[2] SAP Data Quality Management software: http://www.sap.com/pc/tech/enterprise-information-management/software/data-quality/index.html.

[3] Talend Enterprise Data Quality solution: http://www.talend.com/resource/data-quality.html.

5. Repeatedly step 1 to step 4, we eventually accumulate our knowledge on phenomena and data. If we cannot reach an adequate interpretation, then we go back to previous steps.

2.8 Data Analysis

How do we analyse the data which we have in our problem or project? The applied data-centric social sciences are cyber-enabled and require the use of inductive strategies to define problems and solution. One of our final goals in the inductive approach is to find a model to explain the data-generating mechanism. If we have a good model to explain it, we have an ability to predict or to infer the phenomenon which we treat from a subset of the data.

In the data analysis, we can have several types of tools: segmentation, change-point-detection, parameter estimation, classification, correlation, quantification and so on. These methods and tools are addressed in Chap. 3 in detail.

2.9 Data Life-Cycle

Data is often used for a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding is over. Following projects may need data to analyse them or add to the data. The secondary analysis of the data will need the data to analyse them for purposes other than those the primary project intended. As shown in Fig. 2.1, the data creation, recycle and reuse is an ongoing process. The data life-cycle is constructed from the following elements:

- creating data
- processing data
- analysing data
- preserving data
- giving access to data
- reusing data

The data is created from both new data collection which is acquired from the research activity and old data collection which was created at the past activity. Processing data and analysing data are done in the current research project, as we have seen above. The data created in the activity should be preserved as an archive. Giving access to data means that the authors of the data transfer ownership to others and waive the authors' right, which places the work into the public domain. In the European Union, there is the database right. In some countries, there may be no protection for collections of data. When we reuse the data in our own publications, we should indicate the licence under which we are reusing the data in order to make readers to recognise the data reused.

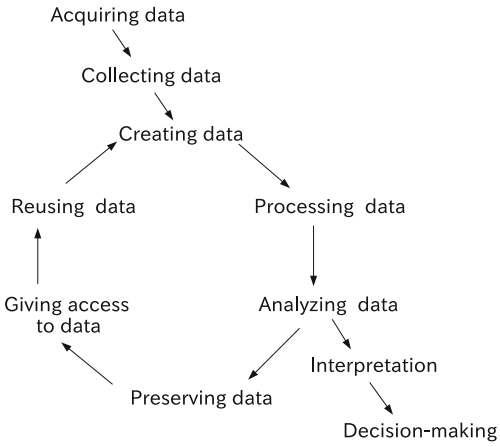


Fig. 2.1 A schematic illustration of data life-cycle

2.10 Social Implementation

2.10.1 Examples

Recently, data-oriented services have launched in several branches of commercial sectors. Facebook [12] and Twitter [35] are currently first examples of social media. ResearchGate [26] and Google Scholar [13] are academic examples. In the case of tourism management systems, Expedia [11], Ebookers [10] and Tripadvisor [32] are good examples. SurveyMonkey [29] enables us to design, collect and analyse our own surveys.

After the Earth Summit, which was held in Rio de Janeiro in 1992, the finiteness of our environment and the importance of monitoring our society was recognised. Several international institutions have issued sustainability indicators in order to guide and facilitate decision-making.

Consequently, social statistics databases from public sectors are available. United Nations Statistical Databases (UNSD)[4] provide sustainable development indicators as well as macro statistics. United Nations Commission on Sustainable Development (CSD) indicators (CSDIs) for Sustainable Development are measured in 14 themes:

- poverty
- natural hazards
- economic development
- governance
- atmosphere
- global economic partnership
- health
- land

[4] http://unstats.un.org/unsd/databases.htm.

- consumption and production patterns
- education
- oceans, seas and coasts
- biodiversity
- demographics
- freshwater

The European Union also selects eleven headline indicators as Sustainable Development Indicators (SDIs)[5]:

- socioeconomic development
- sustainable consumption and production
- social inclusion
- demographic changes
- public health
- climate change and energy
- sustainable transport
- natural resources
- global partnership
- good governance

The World DataBank of the World Bank is a free and open comprehensive data service on socioeconomic-technological systems, which provides several perspectives as macroeconomic indicators related to human activities [31]. Much of the data from the statistical systems of 188 member countries of the International Bank for Reconstruction and Development (IBRD).

Helbing and Balietti proposed the 85 online repositories for the socio-economic sciences [17]. They classified these databases into 18 categories such as:

1. Internet and historical snapshots
2. information retrieval engines
3. text mining on the Web
4. social data sharing
5. conflict data
6. data in economics and finance
7. scientific collaboration data
8. social sciences
9. urban data
10. traffic data
11. open maps
12. logistic data
13. health data
14. climate and environmental data

[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/sdi/indicators.

15. energy
16. reality mining
17. other open data initiatives

These databases define situations of our world from several dimensions such as economy, environment, technology and societies. Some of them have been updated and expanded currently. The current data availability of databases obviously enhance our research and business environment. A future data availability will expand our research to capture our world from data-centric point of view more than the current.

2.10.2 Privacy and Public Utility

2.10.2.1 Data Protection Act

The Data Protection Act 1998 is a United Kingdom Act of Parliament which defines UK law for processing data on identifiable living people [8]. It provides us with the ability to control the area and purpose where our personal information is used by organisations, businesses or the government with the contract at the time when we provide our personal information. Everyone who is responsible for using data has to follow strict rules called *data protection principles*. They must make sure the information is:

- used fairly and lawfully
- used for limited, specifically stated purposes
- used in a way that is adequate, relevant and not excessive
- accurate
- kept for no longer than is absolutely necessary
- handled according to people's data protection rights
- kept safe and secure
- not transferred outside of the company without adequate protection

Fundamentally, usage of personal information other than primary purposes is not permitted. Specifically, ethnic background, political opinions, religious beliefs, health, sexual health and criminal records are sensitive information to be treated with stronger legal protection.

However, the implementation of data protection principles strongly depends on countries. For instance, U.S.-based service providers mostly implement their own privacy policy as self-regulations. The mindset behind this could be summarised as "agree, or stay out". Users who want to use services must accept terms and conditions before they provide their privacy data. European data protection laws are more based on laws and regulations. European privacy protection research has identified three major protection goals of privacy, equivalent to the well-known protection goals of common security such as confidentiality, availability and integrity. In addition to

them, transparency, intervenability and unlinkability are to be considered. Confidentiality is an opposite concept to availability. Transparency is opposite to unlinkability. Integrity is opposite to intervenability. These axes show trade-off relationships in privacy protection.

Demand of secondary usage of data is observed. There is a trade-off relationship between privacy and transparency. Thus, the task of implementing transparency services is a crucial part of all electronic commerce under the regulation.

Anonymisation is one of the key technologies. The data related to privacy is often used for secondary purposes after anonymisation. Several types of anonymising techniques are available. For example, deletion of personal identification numbers and random shuffling are often used. In this case, some methods can keep statistical properties of the data. This technique is called *data anonymisation*. In a data anonymisation process, a real-world application of a privacy-preserving technology, which is called the synthetic data generation, is needed [20]. A plenty of data processing techniques regarding privacy-preserving are recently proposed [1, 22, 33, 36]. The privacy-preserving computations consist of several computations executed in partitioned databases while keeping privacy [1]. The privacy-preserving record linkage techniques [36] allow the linking of databases between organisations while at the same time preserving the privacy of these data. Tsubaki considers a way to evaluate the value of informative data for prediction under partial disclosure [33].

In the study of synthetic data generation [20], there are three types of privacy definition. l-diversity, (d, γ)-privacy and differential privacy. l-diversity can protect against adversaries with background knowledge, but it does not always guarantee privacy when there is a semantic relationship between distinct sensitive values. (d, γ)-privacy is a probabilistic privacy definition in which an adversary believes in some prior probability appearing in the data. Differential privacy is a privacy definition that the anonymisation algorithm should not give additional information about the remaining individual to the adversary who knows complete information about all individuals in the data except one. Jensen also proposes a decentralised solution for supporting an anonymised collection of transparency-relevant information based on the service-oriented principles [19].

In the data validation service, privacy issues are important. Soni et al. propose three types of data validation concepts [28]: producer-centric, customer-centric and reporting-centric. In the provider-centric approach, the actual processing of the data is performed on provider's side, which implies that the relevant data is transferred from consumer to provider. The customer-centric approach shows high privacy but low latency and low efficiency. In the consumer-centric approach, the processing is performed on the consumer side. This results in low privacy but high latency and high efficiency. In the reporting-centric approach, the processing of rules is performed as it is in the consumer-centric approach; however, the flagged data is transferred to the provider for reporting purposes. This shows medium privacy and medium latency but low efficiency.

2.10.3 Problems in Social Implementation

We need to carefully consider social implementation to data-centric social sciences. We eventually recognise several types of problems relating to social implementation of the data-centric approach:

- data lifetime
- data accuracy
- manipulation

The available time of individuals is finite. We often face the problem of data lifetime. Imagine that some data can be shared in some individuals in order to decide their behaviour. The data are created step by step and change gradually. The previous data may mislead the behaviour of the decision-maker. How do we distinguish old data from the latest data? I think that we should observe physical environment and interpret the data with a linkage with the actual environment.

The second problem is data accuracy. The data accuracy should be confirmed based on data from other sources or improved by using several validation procedures. If we found some differences between two databases, then we understand that we need to validate the data from these databases. These data errors may mislead both individual and social behaviour. The manipulation by data is sometimes observed. Some of them are used for the purpose of controlling social behaviour in public sectors or commercial sectors. The Libor (London Interbank Offered Rate) scandal was a series of fraudulent actions. This was that several world's banks obtained profits by manipulating the Libor interest rate illegally [4, 27].

2.10.4 Application of Data Analysis Techniques

Data analysis techniques can be used to detect fraud behaviour. These techniques were firstly employed by banks, telephone companies and insurance companies. The techniques for fraud detection are classified into two main categories including artificial intelligence and statistical techniques [6]. Some of the examples of statistical data analysis techniques are as follows:

- data preprocessing for detecting, validating, correcting error and filling up of incorrect and missing data
- computation of user profile
- matching algorithms for detecting incongruities in the behaviour of users or transactions, which are compared with earlier known profiles or models

To apply these techniques to actual situations, we need to access private data. However, public utility in commercial transactions is sometimes prioritised in comparison with privacy. Other examples are found in drug development. A database concerning medication delivery to the patients has recently been analysed for the purpose of drug

design. Patients' medical data are recorded as electronic health records (EHRs). There are studies on a privacy-protecting information system for controlled disclosure of EHR related to personal data to third parties [16]. The automated healthcare-data-mining system is studied as applications of web technology to healthcare for remote patients [30]. The data-mining service extracts information from data based on a correlation between lifestyle and health data.

References

1. Abbasi, S., Cimato, S., Damiani, E.: Toward secure clustered multi-party computation: a privacy-preserving clustering protocol. In: Mustofa, K., Neuhold, E., Tjoa, A., Weippl, E., You, I. (eds.) Information and Communication Technology. Lecture Notes in Computer Science, vol. 7804, pp. 447–452. Springer, Berlin (2013)
2. AB-ROAD travel retrieval WebAPI: http://webservice.recruit.co.jp/ab-road/
3. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, New York (1998)
4. BBC Timeline: Libor-fixing scandal (6 Feb 2013) http://www.bbc.com/news/business-18671255. Accessed 29 Mar 2014
5. Ben-Gal, I.: Outlier detection. In: Maimon, O., Rockach, L. (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, pp. 131–147. Springer, New York (2005)
6. Chan, P.K., Wei, F., Prodromidis, A.L., Stolfo, S.J.: Distributed data mining in credit card fraud detection. Intel. Syst. Appl. IEEE **14**(6), 67–74 (1999)
7. Chauvenet, W.: A Manual of Spherical and Practical Astronomy V.II, 1st edn. Lippincott, Philadelphia (1863) (Reprint of 1891 5th edn: Dover, NY (1960))
8. Data Protection Act 1998: http://www.legislation.gov.uk/ukpga/1998/29/contents
9. Dean, R.B., Dixon, W.J.: Simplified statistics for small numbers of observations. Anal. Chem. **23**(4), 636–638 (1951)
10. Ebookers: http://www.ebookers.com/
11. Expedia: http://www.expedia.co.jp/
12. Facebook: https://www.facebook.com/
13. Google Scholar: http://scholar.google.co.jp/
14. Google Translate WebAPI: https://developers.google.com/translate/
15. Grubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics **11**(1), 1–21 (1969)
16. Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., Müller, G.: Aspects of privacy for electronic health records. Int. J. Med. Inform. **80**(2), e26–e31 (2011)
17. Helbling, D., Balietti, S.: From social data mining to forecasting socio-economic crises. Eur. Phys. J. Spec. Top. **195**(1), 3–68 (2011)
18. Jalan vacant room information retrieval WebAPI: http://www.jalan.net/jw/jwp0000/jww0001.do
19. Jensen, M.: Towards privacy-friendly transparency services in inter-organizational business processes. In: 2013 IEEE 37th Annual Computer Software and Applications Conference Workshop, pp. 200–205 (2013)
20. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 277–286 (2008)
21. Mahalanobis, P.C.: On the generalised distance in statistics. Proc. Natl. Inst. Sci. India **2**(1), 49–55 (1936)
22. Mehta, S.R., Vinterbo, S.A., Little, S.J.: Ensuring privacy in the study of pathogen genetics. Lancet Infect. Dis. **14**, 70016 (2014)

23. Peirce, B.: Criterion for the rejection of doubtful observations. Astron. J. **2**(43), 161–163 (1852)
24. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Bull. Tech. Comm. Data Eng. **23**(4), 3–13 (2000)
25. Rakuten Web Service WebAPI: http://webservice.rakuten.co.jp/api/ichibaitemsearch/
26. ResearchGate: http://www.researchgate.net/
27. Reuters: Libor scandal may cost banks $14 billion in settlements: analysts (12 July 2012) http://uk.reuters.com/article/2012/07/12/uk-libor-scandal-estimates-idUKBRE86B1EE20120712. Accessed 29 Mar 2014
28. Soni, S., Mehta, S., Hans, S.: Towards providing data validation as a service. In: 2012 IEEE 9th International Conference on Services Computing, pp. 570–577 (2012)
29. SurveyMonkey: https://www.surveymonkey.com/
30. Takeuchi, H., Kodama, N., Hashiguchi, T., Hayashi, D.: Automated healthcare data mining based on a personal dynamic healthcare system. In: Engineering in Medicine and Biology Society, EMBS '06. 28th Annual International Conference of the IEEE, 30 Aug–3 Sept 2006, pp. 3604–3607 (2006)
31. The World DataBank of the World Bank: http://data.worldbank.org
32. Tripadvisor: http://www.tripadvisor.com/
33. Tsubaki, H.: Valuation of partly disclosed datasets for prediction.In: icdmw, 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 733–734 (2013)
34. Tukey, J. W.: Exploratory Data Analysis. Addison-Wesley, Reading (1977)
35. Twitter: https://twitter.com/
36. Vatsalan, D., Christen, P., Verykios, V.S.: A taxonomy of privacy-preserving record linkage techniques. Inform. Syst. **38**(6), 946–969 (2013)
37. Yahoo! JAPAN text analytics WebAPI: http://developer.yahoo.co.jp/webapi/jp/

Part II
Methodology

Chapter 3
Mathematical Expressions

Abstract Statistical methods are useful tools to deal with data on socioeconomic-technological systems. In this chapter, we will address fundamental expressions used in statistics and methods of data analysis: time series analysis, network analysis and spatial analysis.

3.1 Statistical Methods

Uncertainty is deeply related to randomness, which originates from a lack of knowledge. Randomness comes from a large number of possible realisations due to an enormous number of combinations (complexity). Under complexity, we should use a statistical method to understand the meaning of randomness. Knight calls measurable uncertainty risk, and distinguishes it from unmeasurable uncertainty [27].

To measure socioeconomic-technological systems, we can extract information on the underlying characteristics of observations. If the statistical properties are stable over time, then we may use them for our decision making. The measurable uncertainty can be expressed using a probability.

3.1.1 Stochastic Variables and Probability Distributions

Suppose that there is a causality for actual events. Let us consider the value of an output (outcome) X generated from an input (cause) ω, which depends on changes in time and circumstances. Furthermore, suppose that we cannot determine the value of ω for each X from direct observation because of insufficient knowledge or technology.

Even under these circumstances, it is possible to study the values of X that are generated as output repeatedly. In this way, we can gain some understanding of the properties of the output. This, in turn, may make it possible to predict, at least to some degree, what the output will be even in the case of uncertain values of ω.

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_3, © Springer Japan 2014

An observed sequence of outputs, for example, $x_1, x_2, \ldots, x_T$ is called a *sample*. When we express an output X as a function $X(\omega)$ of some uncertain input ω, we consider X to be a stochastic variable.

For stochastic variables, when the value of X is associated with discrete events E_i, which are mutually exclusive ($E_i \bigcap E_j = \emptyset; i \neq j$), the '*probability*' of each state E_i is written as $\Pr[E_i]$.

The probability is the non-negative real-valued function defined by the set of events that fulfils the following principle:

1. The probability of each event occurring is more than 0 and less than 1. Namely, $0 \leq \Pr[E_i] \leq 1$ holds for all i.
2. The probability of the whole event $S = \bigcap_{i=1}^{n} E_i$ occurring is 1: $\Pr[S] = 1$
3. The sum rule holds for countable exclusive events: $\Pr[\bigcup_{i=1}^{n} E_i] = \sum_{i=1}^{n} \Pr[E_i]$

In terms of a consecutive state space, for continuous random variables for which $R = \{-\infty < x < \infty\}$ applies, we can use the probability density function (PDF) $p(x) \geq 0$ to express the probability distribution for a given interval $(a; b]$ as

$$\Pr[a < X \leq b] = \int_a^b p(x)\mathrm{d}x. \tag{3.1}$$

This means that the probability of a stochastic variable X falling within a minute interval $(x; x + \Delta x]$ is approximately equal to the quadrilateral area between the graph $y = p(x)$ and the x-axis ($y = 0$):

$$\Pr[x < X \leq x + \Delta x] \approx p(x)\Delta x. \tag{3.2}$$

In accordance with the definition of a probability, the probability of the whole event is 1. This can also be expressed as a normalisation condition:

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1. \tag{3.3}$$

The probability of X being less than x is expressed as

$$\Pr[X \leq x] = \int_{-\infty}^{x} p(x')\mathrm{d}x'. \tag{3.4}$$

This is called the *cumulative distribution function* (CDF). In addition, the probability of X being greater than x is expressed as

$$\Pr[X > x] = \int_{x}^{\infty} p(x')\mathrm{d}x'. \tag{3.5}$$

This is called the *complementary cumulative distribution function* (CCDF). Clearly, $\Pr[X \leq x] + \Pr[X > x] = 1$ holds.

The CDF $\Pr[X \leq x]$, calculated from Eq. (3.4), is a non-decreasing function. In addition, we can derive a PDF from $\Pr[X \leq x]$ through differentiation owing to the relationship between integration and differentiation.

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} \Pr[X \leq x]. \tag{3.6}$$

Recall that the discrete probability variable X is derived from the discrete state, $R = \{a_1, a_2, a_3, \cdots\}$. Using the concept of the PDF defined in Eq. (3.6), we can express the PDF of the stochastic variable X as

$$p(x) = \sum_{n=1}^{\infty} p_n \delta(x - a_n), \tag{3.7}$$

where $\delta(x)$ is a Dirac δ-function. For a given continuous function $f(x)$ and a given real number x_0, the δ-function fulfils the following characteristics:

$$f(x_0) = \int_{-\infty}^{\infty} f(x)\delta(x - x_0)\mathrm{d}x, \quad \int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1. \tag{3.8}$$

Accordingly, we refer to $p(x)$ as a probability density function in the case of discrete random variables as well as continuous random variables.

The measure of central tendency for stochastic variables are the mode, median and mean. The mode m^* is the value where the probability density function $p(x)$ is maximised:

$$m^* = \arg\max_x p(x). \tag{3.9}$$

The median m_m is the value for which the probability of being higher or lower than X is equal:

$$\Pr[X \leq m_m] = \Pr[X \geq m_m] = \frac{1}{2}. \tag{3.10}$$

The mean m_1 or $\mathrm{E}[X]$ corresponds to the expected value of the stochastic variable X:

$$m_1 = \mathrm{E}[X] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x. \tag{3.11}$$

The mode, median and mean depend on the form of the PDF $p(x)$ and are not necessarily always the same value. We can also use the *variance* defined as

$$\mathrm{Var}[X] = \sigma^2 = \mathrm{E}[(X - m_1)^2] = \int_{-\infty}^{\infty} (x - m_1)^2 p(x) \mathrm{d}x. \tag{3.12}$$

The variance is also calculated as $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. In addition, the square root of the variance is called the standard deviation σ.

Next, we introduce moments as a more generalised approach to the mean and variance.

$$\mathrm{E}[X^r] = \int_{-\infty}^{\infty} x^r p(x) \mathrm{d}x. \tag{3.13}$$

The mean is calculated using $r = 1$, and is the first moment. The variance can be calculated using the $r = 2$ and $r = 1$ moments.

Generally, when the stochastic variable X follows the probability density function $p(x)$, the expected value for the stochastic variable $f(X)$ is calculated as

$$\mathrm{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p(x) \mathrm{d}x. \tag{3.14}$$

In addition to the standard deviation and the three measures of central tendency, other important quantities that characterise the PDF include *skewness* and *kurtosis*. These quantities standardise the standard deviation to the third and fourth power as the third and fourth central moments defined as

$$\lambda_3 = \frac{\mathrm{E}[(X - m_1)^3]}{\sigma^3}, \quad \lambda_4 = \frac{\mathrm{E}[(X - m_1)^4]}{\sigma^4}. \tag{3.15}$$

If the form of the PDF is symmetrical, the skewness is $\lambda_3 = 0$. The kurtosis λ_4 quantifies the peakedness of the PDF.

A normal (Gaussian) distribution is often used. Its PDF is defined as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\Big(-\frac{(x - \mu)^2}{2\sigma_g^2}\Big), \tag{3.16}$$

where μ and σ_g are parameters representing mean and standard deviation, respectively. Equation (3.16) is often denoted as $N(\mu, \sigma_g^2)$. Its CDF and CCDF are calculated as

$$\Pr[X \le x] = \frac{1}{2}\Big[1 + \mathrm{erf}\Big(\frac{x - \mu}{\sqrt{2}\sigma_g}\Big)\Big], \tag{3.17}$$

$$\Pr[X > x] = \frac{1}{2}\left[\mathrm{erfc}\left(\frac{x-\mu}{\sqrt{2}\sigma_g}\right)\right], \tag{3.18}$$

where $\mathrm{erf}(x)$ and $\mathrm{erfc}(x)$ are the error function and the complementary error function. They are, respectively, defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t, \quad \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\mathrm{d}t. \tag{3.19}$$

The skewness and kurtosis of a normal distribution defined in Eq. (3.16) are, respectively, given as $\lambda_3 = 0$ and $\lambda_4 = 3$.

A Poisson distribution is an example of a discrete distribution, which is used in a counting process under a diluted assumption. This is formalised as

$$\Pr[X = n] = \frac{e^{-\lambda_P}\lambda_P^n}{n!}, \tag{3.20}$$

where $\lambda_P(>0)$ is called intensity and n represents a non-negative integer ($n = 0, 1, 2, \ldots$). The mean of Eq. (3.20) is given as $m_1 = \lambda_P$, and its variance as $\sigma^2 = \lambda_P$.

Both the normal and Poisson distributions are derived as different limits of a binomial distribution, respectively.

3.1.2 Sample Moment

The moments corresponding to the stochastic variables and PDFs described above can be estimated using data (samples). Assume that there is a sample $X_i (i = 1, \ldots, T)$ of T items that are independently and randomly selected from the PDF $p(x)$. This is called a random sample. In this case, the simple mean calculated from the sample data is the *sample mean*.

$$\hat{m}_1 = \frac{1}{T}\sum_{i=1}^{T} X_i. \tag{3.21}$$

In addition, the variance calculated from the sample data is the *sample variance*.

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T}(X_i - \hat{m}_1)^2. \tag{3.22}$$

In the same manner, we use the following calculation for the r subsequent sample moments.

$$\hat{m}_r = \frac{1}{T}\sum_{i=1}^{T} X_i^r. \tag{3.23}$$

In accordance with the law of large numbers (described below), as the number of samples increases, the r subsequent sample moments converge to the moments computed from the PDF.

Generally, for a random sample X_i $(i = 1, \ldots, T)$ and $\mathrm{E}[f(X)] < \infty$, the following holds:

$$Y_T = \frac{1}{T}\sum_{i=1}^{T} f(X_i) \overset{a.s.}{\to} \mathrm{E}[f(X)], \tag{3.24}$$

where *a.s.* represents almost sure convergence, such that for a stochastic variable sequence $\{Y_T(\omega)\}$ defined on $\omega \in \Omega$ if and only if ω where $\{Y_T(\omega)\}$ does not converge to the right hand side are included in a zero-probability event.

3.1.3 Major Limit Theorems

Assume that there is an *i.i.d.* stochastic variable series, $X_1, X_2, \ldots$ with a finite mean m_1. The initial T items of the sample mean $S_T = \frac{1}{T}\sum_{i=1}^{T} X_i$ converge to m_1 as $T \to \infty$. This can also be expressed as follows:

$$S_T \overset{a.s.}{\to} m_1, \quad S_T \overset{P}{\to} m_1.$$

The former is called the *strong law of large numbers* and the latter is called the *weak law of large numbers*. Here $\overset{P}{\to}$ represents convergence in probability.

Assume that there is an *i.i.d.* stochastic variable series, $X_1, X_2, \ldots$ with finite mean m_1 and variance σ^2. In this case, the following applies.

$$\sqrt{T}\frac{S_T - m_1}{\sigma} \overset{d}{\to} N(0, 1).$$

This expression describes how the distribution for S_T approaches a normal distribution when T is large. Here, $\overset{d}{\to}$ represents convergence in distribution.

3.1.4 Multivariate Case

The probability of the joint events $X = a_k$, $Y = b_j$ being realised when there are two stochastic variables $X(\omega)$ and $Y(\omega)$ is called *a joint probability distribution*, and

is expressed as

$$\Pr[X = a_k, Y = b_j] = \Pr[a_k, b_j], \quad k, j = 1, 2, \ldots. \tag{3.25}$$

The sum of probabilities over the event is normalised as

$$\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \Pr[a_k, b_j] = 1. \tag{3.26}$$

The probability distribution obtained by the summing a joint probability distribution in terms of one variable with respect to one other variable is called a *marginal probability distribution*:

$$P_X(a_k) = \sum_{j=1}^{\infty} \Pr[a_k, b_j], \tag{3.27}$$

$$P_Y(b_j) = \sum_{k=1}^{\infty} \Pr[a_k, b_j]. \tag{3.28}$$

The joint probability is equivalent to a product of marginal distributions $\Pr[a_k, b_j] = P_X(a_k) P_Y(b_j)$ if and only if two stochastic variables $X(\omega)$ and $Y(\omega)$ are independent of each other.

The joint probability of events $\{X \in A, Y \in B\}$ for two continuous stochastic variables X and Y can be described as

$$\Pr[X \in A, Y \in B] = \int_{x \in A} \int_{y \in B} p(x, y) \mathrm{d}x \mathrm{d}y, \tag{3.29}$$

where $p(x, y)$ is called the *joint probability density function*. The joint probability density function fulfils $p(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \mathrm{d}x \mathrm{d}y = 1$. The marginal probability density function is defined by integrating a joint probability density function in one variable:

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y) \mathrm{d}y, \tag{3.30}$$

$$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) \mathrm{d}x. \tag{3.31}$$

The joint probability density is equivalent to a product of marginal probability density functions, $p(x, y) = p_X(x) p_Y(y)$ if and only if x and y are independent of each other.

Covariance can be defined as

$$\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{3.32}$$

Covariance is a quantity that describes the relationship between the alignment of two stochastic variables. When the covariance is positive, Y has a tendency to increase as X increases, and conversely, to decrease as X decreases. Note that the covariance is a quantity that depends on the scales of X and Y. Therefore, covariance in itself is not very useful. To solve this problem, practical applications often use the standardised correlation coefficient, which is based on the size of the standard deviation and correlation

$$\mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}.$$

The Cauchy-Schwartz inequality indicates that the value of the correlation coefficient satisfies $-1 \leq \mathrm{Corr}[X, Y] \leq 1$. The closer the correlation value is to 1, the more positive the relationship, and the closer it is to -1, the more negative the relationship. If the correlation is 0, then this means that there is no relationship. A value of $\mathrm{Corr}[X, Y] = \pm 1$ indicates perfect correlation.

From observations (x_i, y_i) $(i = 1, \ldots, T)$, covariance $\mathrm{Cov}[X, Y]$ can be approximated as

$$\mathrm{Cov}[X, Y] \approx \frac{1}{T}\sum_{i=1}^{T} x_i y_i - \Big(\frac{1}{T}\sum_{i=1}^{T} x_i\Big)\Big(\frac{1}{T}\sum_{i=1}^{T} y_i\Big), \tag{3.33}$$

which is called the *sample covariance*.

In addition, the *conditional probability density function* can be defined as

$$p(x|y) = \frac{p(x, y)}{p(y)}, \qquad p(y) > 0. \tag{3.34}$$

Using this definition, the *conditional expectation value* can be defined as

$$\mathrm{E}[X|Y] = \int_{-\infty}^{\infty} x p(x|y)\mathrm{d}x. \tag{3.35}$$

For a stochastic variables $X_1, \ldots, X_N$ for which N ($\geq$2) applies, the joint probability density, conditional probability density, and conditional expectation value can be defined in the same manner.

3.1.5 Entropy and Relative Entropy

In this section, we introduce entropy, which is a measure that quantifies the randomness of a phenomenon within the scope of a PDF. The *information entropy* for N-dimensional PDF $p(x_1, \ldots, x_N)$ is given as

$$S(p) = -\int \mathrm{d}x_1 \ldots \int \mathrm{d}x_N \, p(x_1, \ldots, x_N) \ln p(x_1, \ldots, x_N). \tag{3.36}$$

Entropy concerns the ease of predicting event X_i. Now, consider a single variate case ($N = 1$). In the case that the stochastic variable X is often realised around the mode m^*, the entropy becomes a small value. This implies that the mode estimator gives a good prediction.

Relative entropy is the standard for measuring the difference between two PDFs p_1 and p_2. Kullback–Leibler divergence (KL divergence) is one of the most general form of relative entropy. KL divergence is defined as

$$I(p_1, p_2) = \int \mathrm{d}x_1 \ldots \int \mathrm{d}x_N \, p_1(x_1, \ldots, x_N) \ln \frac{p_1(x_1, \ldots, x_N)}{p_2(x_1, \ldots, x_N)}. \tag{3.37}$$

$I(p_1, p_2)$ shows non-negativity. In the inequality $I(p_1, p_2) \geq 0$, the equality is satisfied when $p_1(x_1, \ldots, x_N)$ and $p_2(x_1, \ldots, x_N)$ are identical $p_1 = p_2$ almost everywhere domain of $(x_1, \ldots, x_N)$. This can be derived by using $\ln x \leq x - 1$ (when equality is satisfied with $x = 1$).

$I(p_1, p_2)$ is asymmetric. $I(p_1, p_2) \neq I(p_2, p_1)$ and, accordingly, the triangular inequality $I(p_1, p_2) + I(p_2, p_3) > I(p_1, p_3)$ are not satisfied. As a result, the axiom of distance is not satisfied, which we refer to as quasi-distance. J-divergence, which matches the mutual inversion of relative entropy, $J(p_1, p_2) = I(p_1, p_2) + I(p_2, p_1)$ is often used so that symmetry is satisfied.

Jensen–Shannon divergence is known as another type of relative entropy. Lin proposed a symmetric measure to calculate distance between two probability density functions, called *Jensen–Shannon divergence* (JS divergence), as a new definition for relative entropy [28]. JS divergence allows the probability density function to take a zero value since $0 \ln 0 = 0$, which is proven in Appendix A.

For two density functions, $p_1(x_1, \ldots, x_N)$ and $p_2(x_1, \ldots, x_N)$, JS divergence is defined as

$$JS_2(p_1, p_2) = S\Big(\pi_1 p_1 + \pi_2 p_2\Big) - \pi_1 S(p_1) - \pi_2 S(p_2), \tag{3.38}$$

where $S(p)$ is the information entropy introduced by Eq. (3.36), and the weights π_1 and π_2 must satisfy $\pi_1 + \pi_2 = 1$, for $0 \leq \pi_1 \leq 1$ and $0 \leq \pi_2 \leq 1$.

1. Non-negativity: JS divergence is the non-negative value $JS_2(p_1, p_2) \geq 0$. $JS_2(p_1, p_2) = 0$ is satisfied only when $p_1(x_1, \ldots, x_N) = p_2(x_1, \ldots, x_N)$. almost everywhere domain of $(x_1, \ldots, x_N)$.

2. Symmetry: JS divergence has symmetry $JS_2(p_1, p_2) = JS_2(p_2, p_1)$. However, because JS divergence does not satisfy the triangular inequality either, it is a quasi-distance that does not fully satisfy the distance axiom.

JS divergence can be extended as follows as a measure that for comprehensively quantifying the degree of similarity in the probability density for K items:

$$JS_K(p_1, \ldots, p_K) = S\Big(\sum_{i=1}^{K} \pi_i p_i\Big) - \sum_{i=1}^{K} \pi_i S(p_i), \tag{3.39}$$

where weight π_i must satisfy the characteristic of $\sum_{i=1}^{K} \pi_i = 1$, for $0 \leq \pi_i \leq 1$.

1. Non-negativity: K-variate JS divergence is a non-negative value. $JS_K(p_1, \ldots, p_K) \geq 0$. $JS_K(p_1, \ldots, p_K) = 0$ is only satisfied when $p_1(x_1, \ldots, x_N) = \cdots = p_K(x_1, \ldots, x_N)$ almost everywhere.
2. Convertibility: K-variate JS divergence satisfies convertibility for distributions. $JS_K(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_K) = JS_K(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_K)$.

3.1.6 Maximum Likelihood Estimation

When assuming a model and estimating the model parameters from data, it is natural to select parameters that minimise the difference between the true distribution p_1 assumed from the data and the model distribution p_2. In this case, it would be appropriate to measure the difference between p_1 and p_2 using Kullback–Leibler divergence:

$$\begin{aligned} I(p_1, p_2) &= \int_{-\infty}^{\infty} \mathrm{d}X_1 \ldots \int_{-\infty}^{\infty} \mathrm{d}X_N \, p_1(X_1, \ldots, X_N) \ln\Big(\frac{p_1(X_1, \ldots, X_N)}{p_2(X_1, \ldots, X_N)}\Big) \\ &= \mathrm{E}_{p_1}\Big[\ln\Big(\frac{p_1(X_1, \ldots, X_N)}{p_2(X_1, \ldots, X_N)}\Big)\Big] \\ &= \mathrm{E}_{p_1}[\ln p_1(X_1, \ldots, X_N)] - \mathrm{E}_{p_1}[\ln p_2(X_1, \ldots, X_N)], \end{aligned} \tag{3.40}$$

where $\mathrm{E}_p[X]$ represents the mean of X in terms of the PDF p. Since the first term of the right hand side in Eq. (3.40) only depends on the true PDF of the data, it does not contribute to the minimisation of Eq. (3.40). The second term of the right hand side in Eq. (3.40), which is called a cross entropy to measure the fit between the model and true distributions. The maximisation of the cross entropy $\mathrm{E}_{p_1}[\ln p_2(X_1, \ldots, X_N)]$ implies the minimisation of the distance between p_1 and p_2. Although the true PDF of the data is unknown, if there are a sufficient number of observations, it is possible to compute the cross entropy (the mean logarithmic likelihood) as the sample mean in accordance with the law of large numbers.

In the context of parameter estimation, p_1 is a true PDF generating the data, which is given as T samples $\{x_1, \ldots, x_T\}$, and p_2 is a PDF assumed as a model. If we assume that the sample x_i is sampled from an *i.i.d* distribution, then we employ a simpler model distribution $p_2(X_1, \ldots, X_T) = \prod_{i=1}^{T} p_2(X_i)$. Then, the cross entropy can be approximated as the sample mean in terms of the data,

$$\mathrm{E}_{p_1}[\ln p_2(X_1, \ldots, X_T)] = \mathrm{E}_{p_1}\left[\ln \prod_{i=1}^{T} p_2(X_i)\right] \approx \sum_{i=1}^{T} \ln p_2(x_i). \tag{3.41}$$

Furthermore, we assume a functional form of $p_2(X_1, \ldots, X_T)$ in terms of m parameters $\{\theta_1, \ldots, \theta_m\}$ as $p_2(\boldsymbol{X}; \boldsymbol{\theta})$ and want to estimate adequate parameters $\{\hat{\theta_1}, \ldots, \hat{\theta_m}\}$. Then, the cross entropy is described as a function in terms of the parameters $\{\theta_1, \ldots, \theta_m\}$ and called the *log–likelihood function*:

$$l(\theta_1, \ldots, \theta_m) = \sum_{i=1}^{T} \ln p_2(x_i; \theta_1, \ldots, \theta_m). \tag{3.42}$$

Thus, adequate parameters minimising the Kullback-Leibler divergence defined in Eq. (3.40) are given by maximising the cross entropy $\mathrm{E}_{p_1}[\ln p_2(X_1, \ldots, X_N)]$:

$$\{\hat{\theta_1}, \ldots, \hat{\theta_m}\} = \arg \max_{\theta_1, \ldots, \theta_m} l(\theta_1, \ldots, \theta_m). \tag{3.43}$$

This solution can be obtained from likelihood equations

$$\frac{\partial l}{\partial \theta_1} = \cdots = \frac{\partial l}{\partial \theta_m} = 0. \tag{3.44}$$

The Cramér-Rao inequality provides us with a lower bound on the variance-covariance matrix of the bias $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. The *score* of the observations $\{x_1, \ldots, x_T\}$ for the assumed PDF $p_2(x_i; \theta_1, \ldots, \theta_m)$ is defined as

$$Y_{ki}(\theta_1, \ldots, \theta_m) = \frac{\partial}{\partial \theta_k} \ln p_2(x_i; \theta_1, \ldots, \theta_m). \tag{3.45}$$

The *Fisher information* is defined as the variance of the scores

$$F_{ij} = \mathrm{E}\Big[\frac{\partial}{\partial \theta_i} \ln p_2(x; \theta_1, \ldots, \theta_m) \frac{\partial}{\partial \theta_j} \ln p_2(x; \theta_1, \ldots, \theta_m)\Big], \tag{3.46}$$

or the second partial derivatives of the Shannon entropy of p_2 in terms of parameters $\theta_i \quad (i = 1, \ldots, m)$.

$$F_{ij} = -\mathrm{E}\Big[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_2(x; \theta_1, \ldots, \theta_m)\Big], \tag{3.47}$$

which it can be derived as the Hessian of the log–likelihood function. This gives the lower bound of variance-covariance matrix of the bias:

$$\mathrm{Cov}[\hat{\theta}_i - \theta_i, \hat{\theta}_j - \theta_j] \geq [\boldsymbol{F}^{-1}]_{ij}. \tag{3.48}$$

Moreover, if we assume that a model is Markovian, then we may employ a model distribution $p_2(X_1, \ldots, X_T) = \prod_{i=2}^{T} q_2(X_i|X_{i-1}; \theta_{m+1}, \ldots, \theta_{m+n}) r_2(X_1; \theta_1, \ldots, \theta_m)$, where $q_2(x_2|x_1; \theta_{m+1}, \ldots, \theta_{m+n})$ is a PDF of x_2 conditioning on x_1 and $r_2(X_1; \theta_1, \ldots, \theta_m)$ is an unconditional PDF of X. In this case, the log–likelihood function is written as

$$l(\theta_1, \ldots, \theta_{m+n}) = \sum_{i=2}^{T} \ln q_2(X_i|X_{i-1}; \theta_{m+1}, \ldots, \theta_{m+n}) + \ln r_2(X_1; \theta_1, \ldots, \theta_m). \tag{3.49}$$

However, we cannot always get the solution of Eq. (3.43) in an analytical manner. We often need to use a numerical method to solve it. To solve this optimisation problem, we use a gradient method. Furthermore, the log–likelihood function is not always unimodal and its convexity is not always guaranteed. Therefore, we need to calculate optimised parameters from different initial parameter values in several trials, and then we choose the most optimal ones as parameter estimates. Note that this method does not always guarantee the global optima of Eq. (3.43).

3.1.7 Gradient Method

In most cases, Eq. (3.43) cannot be solved in any analytical manner. Thus, we must use numerical optimisation methods. In fact, there are many methods to find an optimal solution numerically. Gradient ascent (descent) is the first-order optimisation algorithm to find a local maximum (minimum) of a function using gradient ascent (decent). Suppose that $U(\boldsymbol{x})$ is a scalar function in terms of m-dimensional multivariate variable $\boldsymbol{x}$. $U(\boldsymbol{x})$ increases (decreases) fastest if one goes from $\boldsymbol{a}$ in the direction of the positive (negative) gradient at $\boldsymbol{a}$, $\nabla U(\boldsymbol{a}) = \sum_{i=1}^{m} \frac{\partial U}{\partial x_i}(\boldsymbol{a}) \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ represents the i-th orthogonal unit vector in the m-dimensional space, such that $\boldsymbol{e}_i \cdot \boldsymbol{e}_j = \delta_{ij}$ for $i, j = 1, \ldots, m$. For a small value γ_n, the sequence $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ such that

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n \pm \gamma_n \nabla U(\boldsymbol{x}_n), \tag{3.50}$$

may converge to the desired local maximum (minimum) of $U(\boldsymbol{x})$ corresponding to an initial value $\boldsymbol{x}_0$. The conjugate gradient (CG) method is an algorithm for the numerical solution of particular systems of linear equations and is often implemented as an iterative algorithm. The stochastic gradient descent (SGD) is recently proposed to solve an optimisation problem in machine learning [14].

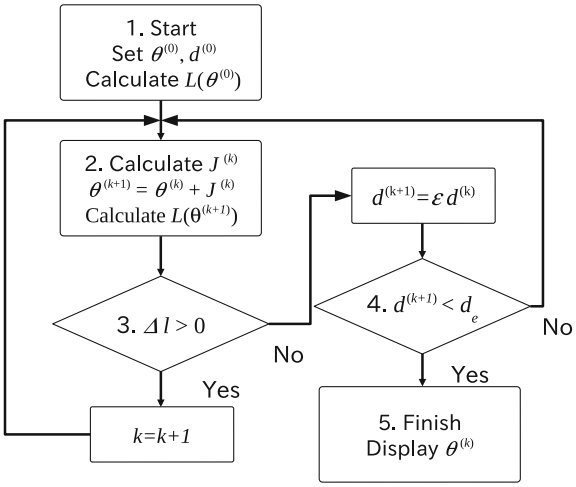


Fig. 3.1 Flow chart of the gradient method for maximisation

Here, we will see a simple local optimal search algorithm with the local gradient to understand the mechanism. For a sake of simplicity, let us consider an algorithm to find adequate parameters maximising the likelihood function with three parameters ($m = 3$) $l(\theta_1, \theta_2, \theta_3)$. Figure 3.1 shows a flow chart of the gradient method.

1. Set initial values $(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})$ and step size $d^{(0)}$.
2. Calculate $J_i^{(k)} (i = 1, 2, 3)$ as

$$\frac{\partial l}{\partial \theta_1} \approx J_1{}^{(k)} = \frac{l(\theta_1^{(k)} + \Delta\theta_1, \theta_2^{(k)}, \theta_3^{(k)}) - l(\theta_1^{(k)} - \Delta\theta_1, \theta_2^{(k)}, \theta_3^{(k)})}{2\Delta\theta_1},$$
$$\frac{\partial l}{\partial \theta_2} \approx J_2{}^{(k)} = \frac{l(\theta_1^{(k)}, \theta_2^{(k)} + \Delta\theta_2, \theta_3^{(k)}) - l(\theta_1^{(k)}, \theta_2^{(k)} - \Delta\theta_2, \theta_3^{(k)})}{2\Delta\theta_2},$$
$$\frac{\partial l}{\partial \theta_3} \approx J_3{}^{(k)} = \frac{l(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)} + \Delta\theta_3) - l(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)} - \Delta\theta_3)}{2\Delta\theta_3},$$

where k indicates time step. Update the parameters as $\theta_i^{(k+1)} = \theta_i^{(k)} + d_i^{(k)} (i = 1, 2, 3)$, in which $d_i^{(k)}$ is determined by

$$d_i^{(k)} = d^{(k)} \frac{J_i^{(k)}}{\sqrt{\left(J_1^{(k)}\right)^2 + \left(J_2^{(k)}\right)^2 + \left(J_3^{(k)}\right)^2}} \quad (i = 1, 2, 3).$$

3. If $\Delta l = l(\theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3^{(k+1)}) - l(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$ becomes more than 0, update the step size $d^{(k+1)} = d^{(k)}$ and then go to step 2. If not, update the step size as $d^{(k+1)} = \varepsilon d^{(k)}$ $(0 < \varepsilon < 1)$ and then go to step 4 (e.g. $\varepsilon = 0.1$).

4. If $d^{(k)} < d_e$, then go to step 5. If not, then go to step 2. This is a terminate condition, where d_e is a positive constant (e.g. $d_e = 10^{-10}$).
5. Stop the program and display $(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$ as parameter estimates.

Thus, we obtain $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ such that $J_i = 0\,(i = 1, 2, 3)$ and $l(\theta_1, \theta_2, \theta_3)$ is locally maximised. For given data that are assumed to obey a certain distribution, we set the log–likelihood function $l(\theta_1, \theta_2, \theta_3)$ and control parameters $(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})$, $d^{(0)}$, $\Delta\theta_1$, $\Delta\theta_2$, $\Delta\theta_3$, ε, and d_e. Then, we update $(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$ to maximise $l(\theta_1, \theta_2, \theta_3)$.

3.1.8 Information Criteria

Information criteria are used to evaluate which models are better than other. Generally, the higher number of parameters K in the model indicates that the maximum log–likelihood value tends to be higher. However, it is believed that models in which the number of parameters is too high contain more error in parameter estimates than the model with the less number of parameters.

The Akaike information criterion is an information criterion that considers how well the data fits the model and penalises for higher numbers of parameters [1]. The *Akaike information criterion* (AIC) can be defined as

$$AIC = -2 \times (\text{maximum log–likelihood value}) + 2 \times (\text{freedom of parameters}). \tag{3.51}$$

Another frequently used information criterion is called the *Bayesian information criterion*. There are several definitions for the Bayesian information criterion (BIC), one of which can be written as follows:

$$BIC = -2 \times (\text{maximum log–likelihood value}) + (\text{freedom of parameters}) \times \ln(\text{data length}). \tag{3.52}$$

BIC imposes penalties towards for parameters that change depending on the data length.

3.1.9 Regression Analysis

Suppose that we have T sets of some variables and that we want to examine a dependence among the variables or to explain a variable with other variables. Then, regression analysis is required.

There are several types of regression methods such as *ordinary least squared regression* (OLS) [44], generalised least squares [26], *reduced major axis regression* (RMA) [23, 41] and *major axis regression* (MAR). In this subsection, we will address two regression methods, OLS and RMA regressions. The underlying idea of regression analysis is to find the parameters of a function while minimising errors between the curve and the data.

3.1.9.1 An Ordinary Least Squared Regression

Let us start with a simple case. Suppose an OLS regression of Y on X for T observations (x_i, y_i) $(i = 1, \ldots, T)$ under the assumption that

$$Y = aX + b + e. \tag{3.53}$$

Y is called the *explained variable* and X the *explanatory variable*. All variation that is not explained by the line $Y = aX + b$ is expressed as error term e in Eq. (3.53). The value of e for each subject is identical to the residual for that subject. Then, we want to find the best estimate of model parameters a and b that minimises the sum of squared residuals.

$$\{\hat{a}, \hat{b}\} = \min_{a,b}\Big[\sum_{i=1}^{T}\Big(y_i - ax_i - b\Big)^2\Big]. \tag{3.54}$$

Partially differentiating Eq. (3.54) in terms of parameters a and b and setting them into zero, we obtain

$$\begin{bmatrix} T & \sum_{i=1}^{T} x_i \\ \sum_{i=1}^{T} x_i & \sum_{i=1}^{T} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{T} y_i \\ \sum_{i=1}^{T} x_i y_i \end{bmatrix}, \tag{3.55}$$

which is called the *normal equations*. Equation (3.55) can be solved as follows:

$$\begin{aligned} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} &= \begin{bmatrix} T & \sum_{i=1}^{T} x_i \\ \sum_{i=1}^{T} x_i & \sum_{i=1}^{T} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{T} y_i \\ \sum_{i=1}^{T} x_i y_i \end{bmatrix} \\ &= \frac{1}{T\sum_{i=1}^{T} x_i^2 - \Big(\sum_{i=1}^{T} x_i\Big)^2} \begin{bmatrix} \sum_{i=1}^{T} x_i^2 & -\sum_{i=1}^{T} x_i \\ -\sum_{i=1}^{T} x_i & T \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{T} y_i \\ \sum_{i=1}^{T} x_i y_i \end{bmatrix}. \end{aligned} \tag{3.56}$$

Thus, the OLS slope a of Y on X is given as

$$\hat{a} = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}, \tag{3.57}$$

and the intercept is described as

$$\hat{b} = \frac{\mathrm{E}[X^2]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[XY]}{\mathrm{Var}[X]} = \mathrm{E}[Y] - \hat{a}\mathrm{E}[X]. \tag{3.58}$$

Errors of the parameter estimates can be calculated as [44]

$$\sigma_a = \sqrt{\frac{MSE}{T\mathrm{Var}[X]}}, \tag{3.59}$$

$$\sigma_b = \sqrt{MSE\left(\frac{1}{T} + \frac{\mathrm{E}[X]^2}{T\mathrm{Var}[X]}\right)}, \tag{3.60}$$

where the mean square error MSE is described as

$$MSE = \frac{1}{T-2}\sum_{i=1}^{T}\left(y_i - \hat{a}x_i - \hat{b}\right)^2 = \left(\mathrm{Var}[Y] - \frac{\mathrm{Cov}[X, Y]^2}{\mathrm{Var}[X]}\right)\frac{T}{T-2}. \tag{3.61}$$

Equation (3.61) is derived as follows: From Eq. (3.58), the mean square error MSE can be written as

$$\begin{aligned} MSE &= \frac{1}{T-2}\sum_{i=1}^{T}\left(y_i - \mathrm{E}[Y] - \hat{a}(x_i - \mathrm{E}[X])\right)^2 \\ &= \frac{1}{T-2}\sum_{i=1}^{T}\left\{(y_i - \mathrm{E}[Y])^2 + \hat{a}^2(x_i - \mathrm{E}[X])^2 - 2\hat{a}(x_i - \mathrm{E}[X])(y_i - \mathrm{E}[Y])\right\} \\ &= \left\{\mathrm{Var}[Y] + \hat{a}^2\mathrm{Var}[X] - 2\hat{a}\mathrm{Cov}[X, Y]\right\}\frac{T}{T-2}. \end{aligned} \tag{3.62}$$

Inserting Eq. (3.57) into Eq. (3.62), we have

$$MSE = \left(\mathrm{Var}[Y] - \frac{\mathrm{Cov}[X, Y]^2}{\mathrm{Var}[X]}\right)\frac{T}{T-2}. \tag{3.63}$$

3.1.9.2 A Reduced Major Axis Regression

A regression analysis assumes that the equation allows for error in both X and Y:

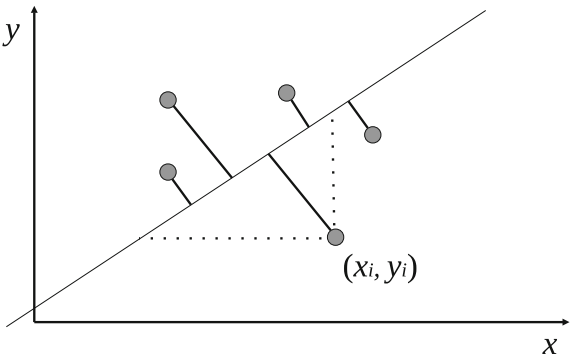


Fig. 3.2 A schematic illustration of a reduced major axis regression

$$Y + u = aX + b + e, \tag{3.64}$$

where u represents errors in the observed values of X. If X is measured with error, then the X-axis trait has an added component of variation and should be expressed as $(X + u)$. The slope becomes

$$a_{observed} = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X] + \mathrm{Var}[u]}. \tag{3.65}$$

The error in X causes the slope of Eq. (3.65) to be shallower since it has a larger denominator than the OLS slope calculated if X is measured without error in Eq. (3.57). The bias in the slope calculated with Eq. (3.65) is known as attenuation or regression dilution. Thus, error in X and error in Y have different consequences for an OLS regression. Another problem is called the symmetry problem. This second source of concern about OLS is not directly related to the errors but is based on the lack of symmetry between the OLS regression of Y on X and of X on Y.

To improve these drawbacks of the OLS regression, an RMA regression is considered [40]. We assume a linear relationship between the explanatory variable X and the explained variable Y:

$$Y = aX + b, \tag{3.66}$$

where a and b are a slope and intercept. Let us find an adequate line with a and b by minimising the sum of the area of the triangles constructed from the line and a data point. Figure 3.2 shows a conceptual illustration of the RMA regression. Consider the area of a triangle of the line $(y = ax + b)$ and the i-th data point (x_i, y_i). The area of this triangle is calculated as

$$\frac{1}{2}\left|x_i - \frac{y_i - b}{a}\right|\left|ax_i + b - y_i\right| = \frac{1}{2}\frac{(ax_i + b - y_i)^2}{|a|}.$$

Therefore, the total area of the triangles computed from T data points, which is an objective function, is calculated as

$$f(a,b) = \frac{1}{2}\sum_{i=1}^{T}\frac{(ax_i + b - y_i)^2}{|a|}. \tag{3.67}$$

For $a > 0$, minimising $f(a,b)$ in terms of a and b implies

$$\frac{\partial f}{\partial a} = \frac{1}{2a^2}\sum_{i=1}^{T}[a \cdot 2(ax_i + b - y_i)x_i - (ax_i + b - y_i)^2] = 0, \tag{3.68}$$

$$\frac{\partial f}{\partial b} = \frac{1}{2}\sum_{i=1}^{T}\frac{2(ax_i + b - y_i)}{a} = 0. \tag{3.69}$$

From Eq. (3.69), we have

$$\hat{b} = \frac{\sum_{i=1}^{T} y_i}{T} - \hat{a}\frac{\sum_{i=1}^{T} x_i}{T}. \tag{3.70}$$

Inserting Eq. (3.70) into Eq. (3.68), we obtain

$$\hat{a}^2\Big[\sum_{i=1}^{T} x_i^2 - \frac{(\sum_{i=1}^{T} x_i)^2}{T}\Big] - \Big[\sum_{i=1}^{T} y_i^2 - \frac{(\sum_{i=1}^{T} y_i)^2}{T}\Big] = 0. \tag{3.71}$$

Thus, since we impose $a > 0$, we have

$$\hat{a} = \sqrt{\frac{\sum_{i=1}^{T} y_i^2 - \frac{(\sum_{i=1}^{T} y_i)^2}{T}}{\sum_{i=1}^{T} x_i^2 - \frac{(\sum_{i=1}^{T} x_i)^2}{T}}}. \tag{3.72}$$

For $a < 0$, we obtain the same equations as Eqs. (3.70) and (3.72) from Eqs. (3.68) and (3.69). Consequently, since we impose $a < 0$, we get

$$\hat{a} = -\sqrt{\frac{\sum_{i=1}^{T} y_i^2 - \frac{(\sum_{i=1}^{T} y_i)^2}{T}}{\sum_{i=1}^{T} x_i^2 - \frac{(\sum_{i=1}^{T} x_i)^2}{T}}}. \tag{3.73}$$

Equations (3.72) and (3.73) are also expressed as

$$\hat{a} = \pm\sqrt{\frac{\mathrm{Var}[Y]}{\mathrm{Var}[X]}}, \tag{3.74}$$

Therefore, the sign of $\hat{a}$ should be chosen according to the sign of the second derivative of $f(a, b)$ in terms of a. Namely, the sign of $\hat{a}$ should be determined as a value satisfying $\frac{\partial^2 f}{\partial a^2} < 0$. Thus, the sign of $\hat{a}$ is equivalent to the sign of $\mathrm{Cov}[X, Y]$. We can also obtain $\hat{b}$ from Eq. (3.70) and Eq. (3.74). The coefficient of determination $r^2 = \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]\mathrm{Var}[Y]}$ and errors are calculated as

$$\sigma_a = \sqrt{\frac{MSE}{T\mathrm{Var}[X]}}, \tag{3.75}$$

$$\sigma_b = \sqrt{MSE\left(\frac{1}{T} + \frac{\mathrm{E}[X]^2}{T\mathrm{Var}[X]}\right)}, \tag{3.76}$$

where the mean square error MSE is computed as

$$MSE = \frac{1}{T-2}\sum_{i=1}^{T}(y_i - \hat{a}x_i - \hat{b})^2 = \Big(\mathrm{Var}[Y] - \hat{a}\mathrm{Cov}[X, Y]\Big)\frac{2T}{T-2}, \tag{3.77}$$

which is derived in Appendix B.

As discussed by Ricker [37], the difference between the two OLS lines (of Y on X and of X on Y) for a single data set is normally observed. This allows the two OLS regressions to be examined on a single set of axes. The two lines will intercept at $\mathrm{E}[X]$ and $\mathrm{E}[Y]$.

Figure 3.3 shows the difference between the two OLS lines (of Y on X and of X on Y) for a single data set. The dashed lines express the two OLS lines. The solid line represents the RMA line. Three fitting lines are computed for the same dataset (the common logarithm of GDP per capita and the common logarithm of CO_2 emissions per capita in 2000). Specifically, when we compute coefficients of an allometric relationship, we should prefer the RMA regression to the OLS regression because of its symmetricity. Table 3.1 shows parameter estimates with the three cases. The slopes estimated from the three methods are slightly different from one another.

3.1.9.3 Alternative Derivation of RMA Regression Using Kullback–Leibler Divergence

Let us derive the same coefficients as the RMA regression, given as Eqs. (3.70) and (3.74) from an alternative perspective.

A fundamental idea of this derivation is equivalence between the marginal distribution in terms of x and a distribution of x transformed from the marginal distribution in terms of y under an assumed function or vice versa.

We assume that we have T data points of (x_i, y_i). Let $p_Y(y)$ be denoted as a marginal distribution of y, and $p_X(x)$ as a marginal distribution of x. Namely, the

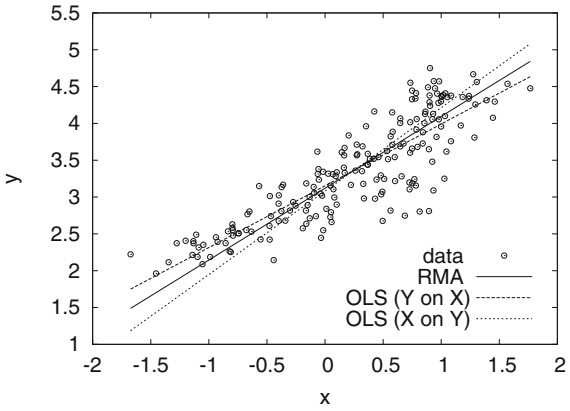


Fig. 3.3 Scatter diagram of the data showing the RMA regression and OLS regression of Y on X and of X on Y. The X represents common logarithms of GDP per capita and the Y common logarithms of CO_2 emissions per capita for 188 countries and territories ($T = 188$) in 2000. The data source is DataBank of the World Bank

Table 3.1 Parameter estimates of regression analysis for GDP per capita in 2000 (X) and CO_2 emissions per capita in 2000 (Y)

Method	$\hat{a}$	$\hat{b}$
OLS (X on Y)	1.13269	3.08109
OLS (Y on X)	0.83864	3.15315
RMA	0.97464	3.11982

observations x_i are assumed to be sampled from $p_X(x)$ and the observations y_i from $p_Y(y)$.

Consider a marginal probability density function $q_X(x)$ that is derived from the transformation of a stochastic variable y into x using a linear function $x = (y-b)/a$. Using a transformation formula of the stochastic variable $q_X(x)\mathrm{d}x = p_Y(y)\mathrm{d}y$, we have

$$q_X(x) = p_Y(ax+b)|a|. \tag{3.78}$$

Next, let us measure a distance between $q_X(x)$ and $p_X(x)$ using Kullback–Leibler divergence:

$$KL(p_X, q_X) = \int_{-\infty}^{\infty} p_X(x) \ln p_X(x)\mathrm{d}x - \int_{-\infty}^{\infty} p_X(x) \ln p_Y(ax+b)\mathrm{d}x - \ln|a|. \tag{3.79}$$

The maximisation or minimisation of Eq. (3.79) can be given by $\frac{\partial KL}{\partial a} = 0$ and $\frac{\partial KL}{\partial b} = 0$. Namely, we have

$$\int_{-\infty}^{\infty} p_X(x) \frac{p'_Y(ax+b)x}{p_Y(ax+b)} \mathrm{d}x = -\frac{1}{a}, \tag{3.80}$$

$$\int_{-\infty}^{\infty} p_X(x) \frac{p'_Y(ax+b)}{p_Y(ax+b)} \mathrm{d}x = 0, \tag{3.81}$$

where $p'(x)$ denotes $\frac{\mathrm{d}p}{\mathrm{d}x}$. In contrast, consider a marginal probability density function $q_Y(y)$ that is derived from the transformation of a stochastic variable x into y using the linear function $y = ax + b$.

$$q_Y(y) = p_X\left(\frac{y-b}{a}\right)\frac{1}{|a|}. \tag{3.82}$$

Here, let us also measure the distance between $q_Y(y)$ and $p_Y(y)$ using Kullback–Leibler divergence.

$$KL(p_Y, q_Y) = \int_{-\infty}^{\infty} p_Y(y) \ln p_Y(y) \mathrm{d}y - \int_{-\infty}^{\infty} p_Y(y) \ln p_X\left(\frac{y-b}{a}\right) \mathrm{d}y + \ln |a|. \tag{3.83}$$

The maximisation or minimisation of Eq. (3.83) can be derived from $\frac{\partial KL}{\partial a} = 0$ and $\frac{\partial KL}{\partial b} = 0$. Thus, we have

$$\int_{-\infty}^{\infty} p_Y(y) \frac{p'_X\left(\frac{y-b}{a}\right)}{p_X\left(\frac{y-b}{a}\right)} (y-b) \mathrm{d}y = a, \tag{3.84}$$

$$\frac{1}{a} \int_{-\infty}^{\infty} p_Y(y) \frac{p'_X\left(\frac{y-b}{a}\right)}{p_X\left(\frac{y-b}{a}\right)} \mathrm{d}y = 0. \tag{3.85}$$

Therefore, we have

$$a = \int_{-\infty}^{\infty} p_Y(y) \frac{p'_X\left(\frac{y-b}{a}\right) y}{p_X\left(\frac{y-b}{a}\right)} \mathrm{d}y, \tag{3.86}$$

$$\int_{-\infty}^{\infty} p_Y(y) \frac{p'_X\left(\frac{y-b}{a}\right)}{p_X\left(\frac{y-b}{a}\right)} \mathrm{d}y = 0. \tag{3.87}$$

Suppose that $p_X(x)$ and $p_Y(y)$, respectively, are normal distributions parametrised by μ_X, μ_Y, σ_X^2 and σ_Y^2:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\Big[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\Big], \tag{3.88}$$

$$p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\Big[-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\Big]. \tag{3.89}$$

Then, since from Eq. (3.89) we have the equality

$$\frac{p_Y'(ax+b)}{p_Y(ax+b)} = -\frac{ax+b-\mu_Y}{\sigma_Y^2}, \tag{3.90}$$

inserting it into Eq. (3.80), we get

$$\begin{aligned}
\frac{1}{a} &= \frac{a\mathrm{E}[X^2] + b\mathrm{E}[X] - \mu_Y \mathrm{E}[X]}{\sigma_Y^2}, \\
\frac{1}{a} &= \frac{a\mathrm{E}[X^2] + b\mathrm{E}[X] - \mathrm{E}[Y]\mathrm{E}[X]}{\mathrm{Var}[Y]}, \\
\mathrm{E}[X^2]a^2 + ab\mathrm{E}[X] - a\mathrm{E}[X]\mathrm{E}[Y] &= \mathrm{Var}[Y].
\end{aligned} \tag{3.91}$$

From Eq. (3.81), we have

$$\begin{aligned}
a\mathrm{E}[X] + b &= \mathrm{E}[Y], \\
b &= \mathrm{E}[Y] - a\mathrm{E}[X].
\end{aligned} \tag{3.92}$$

Inserting Eq. (3.92) into Eq. (3.91), we obtain

$$\begin{aligned}
\mathrm{E}[X^2]a^2 - \mathrm{E}[X]^2 a^2 &= \mathrm{Var}[Y], \\
\mathrm{Var}[X]a^2 &= \mathrm{Var}[Y].
\end{aligned} \tag{3.93}$$

Equations (3.92) and (3.93) imply that the adequate slope $\hat{a}$ and intercept $\hat{b}$ are given as

$$\hat{a} = \pm\sqrt{\frac{\mathrm{Var}[Y]}{\mathrm{Var}[X]}}, \tag{3.94}$$

$$\hat{b} = \mathrm{E}[Y] - \hat{a}\mathrm{E}[X]. \tag{3.95}$$

Next, let us consider the case of X. Since from Eq. (3.88) we have the equality

$$\frac{p'_X\left(\frac{y-b}{a}\right)}{p_X\left(\frac{y-b}{a}\right)} = -\frac{(y-b)/a - \mu_X}{\sigma_X^2}, \tag{3.96}$$

Equation (3.86) can be rewritten as

$$\begin{aligned} a &= \frac{\frac{1}{a}\mathrm{E}[Y^2] + \frac{b}{a}\mathrm{E}[Y] - \mu_X \mathrm{E}[Y]}{\sigma_X^2}, \\ \mathrm{Var}[X]a^2 &= \mathrm{E}[Y^2] + b\mathrm{E}[Y] - \mathrm{E}[Y]\mathrm{E}[X]a, \\ \mathrm{Var}[X]a^2 &- \mathrm{E}[Y^2] - b\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y]a = 0. \end{aligned} \tag{3.97}$$

From Eq. (3.85), we get

$$\frac{\mathrm{E}[Y] - b}{a\mathrm{Var}[X]} = \frac{\mathrm{E}[X]}{\mathrm{Var}[X]}. \tag{3.98}$$

Namely, we obtain

$$b = \mathrm{E}[Y] - a\mathrm{E}[X]. \tag{3.99}$$

Inserting Eq. (3.99) into Eq. (3.97), we obtain

$$\mathrm{Var}[X]a^2 = \mathrm{Var}[Y]. \tag{3.100}$$

Namely, Eqs. (3.99) and (3.100) imply

$$\hat{a} = \pm\sqrt{\frac{\mathrm{Var}[Y]}{\mathrm{Var}[X]}}, \tag{3.101}$$

$$\hat{b} = \mathrm{E}[Y] - \hat{a}\mathrm{E}[X]. \tag{3.102}$$

Thus, Eqs. (3.94) and (3.101) are the same equations as each other. The sign of a can be determined by considering the sign of the second-order derivative of $KL(p_X, q_X)$ in terms of a. a is given as a value satisfying $\frac{\partial^2 KL}{\partial a^2} < 0$. Thus, the sign of $\hat{a}$ is equivalent to the sign of $\mathrm{Cov}[X, Y]$.

3.1.9.4 OLS Regression Derived from Maximum Likelihood Estimation

Suppose that we have the data sequence (x_i, y_i) $(i = 1, \ldots, T)$. We also assume the following model parametrised by $a_0, \ldots, a_m$:

$$y = f(x) = a_0 + a_1 x + \cdots + a_m x^m. \tag{3.103}$$

From Eq. (3.103), we introduce an error u_i between $f(x_i)$ and y_i,

$$u_i = y_i - f(x_i) = y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m. \tag{3.104}$$

Then, we want to find the best estimate of model parameters $a_0, \ldots, a_m$ in the sense of minimising the sum of squared residuals.

$$\{\hat{a}_0, \ldots, \hat{a}_m\} = \min_{a_0,\ldots,a_m} \Big[\sum_{i=1}^{T} \Big(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m \Big)^2 \Big]. \tag{3.105}$$

Partially differentiating the sum of squared residuals in Eq. (3.105) in terms of parameters $a_0, a_1, \ldots, a_m$ and setting them into zero, we obtain the *normal equations*

$$\begin{bmatrix} T & \sum_{i=1}^{T} x_i & \sum_{i=1}^{T} x_i^2 & \cdots & \sum_{i=1}^{T} x_i^m \\ \sum_{i=1}^{T} x_i & \sum_{i=1}^{T} x_i^2 & \sum_{i=1}^{T} x_i^3 & \cdots & \sum_{i=1}^{T} x_i^{m+1} \\ \vdots & \vdots & \ddots & & \vdots \\ \sum_{i=1}^{T} x_i^m & \sum_{i=1}^{T} x_i^{m+1} & \sum_{i=1}^{T} x_i^{m+2} & \cdots & \sum_{i=1}^{T} x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{T} y_i \\ \sum_{i=1}^{T} x_i y_i \\ \vdots \\ \sum_{i=1}^{T} x_i^m y_i \end{bmatrix}. \tag{3.106}$$

We shall derive the same equations from a different perspective. Assume that u_i are sampled from *i.i.d.* normal distributions with zero mean and variance σ_u^2:

$$p(y|\mu(x; a_0, a_1, \ldots, a_m), \sigma_u^2) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\Big[-\frac{(y - (a_0 + a_1 x + \cdots + a_m x^m))^2}{2\sigma_u^2} \Big]. \tag{3.107}$$

Then, the log–likelihood function can be written as

$$\begin{aligned} l(a_0, a_1, \ldots, a_m, \sigma_u^2) &= \sum_{i=1}^{T} \ln p(y_i|\mu(x_i; a_0, a_1, \ldots, a_m), \sigma_u^2) \\ &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln \sigma_u^2 - \frac{1}{2\sigma_u^2} \sum_{i=1}^{T} \Big(y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m) \Big)^2. \end{aligned} \tag{3.108}$$

Therefore, partially differentiating Eq. (3.108) in terms of a_i ($i = 0, 1, \ldots, m$) and setting them at zero imply

$$\frac{\partial l}{\partial a_0} = \frac{1}{\sigma_u^2} \sum_{i=1}^{T} \Big(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m \Big) = 0,$$

$$\frac{\partial l}{\partial a_1} = \frac{1}{\sigma_u^2}\sum_{i=1}^{T} x_i\left(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m\right) = 0,$$

$$\vdots$$

$$\frac{\partial l}{\partial a_m} = \frac{1}{\sigma_u^2}\sum_{i=1}^{T} x_i^m\left(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m\right) = 0.$$

Thus, we get the same equations as Eq. (3.106). Moreover, a partial derivative in terms of σ_u^2 gives

$$-\frac{T}{2\sigma_u^2} + \frac{1}{2\sigma_u^4}\sum_{i=1}^{T}(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2 = 0. \tag{3.109}$$

Thus, we have

$$\hat{\sigma}_u^2 = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{a}_0 - \hat{a}_1 x_i - \cdots - \hat{a}_m x_i^m)^2. \tag{3.110}$$

As a result, the maximum log–likelihood value is obtained as

$$l(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_m, \hat{\sigma}_u^2) = -\frac{T}{2}(\ln(2\pi) + 1 + \ln\hat{\sigma}_u^2). \tag{3.111}$$

Therefore, AIC and BIC are written as

$$\mathrm{AIC} = T(\ln(2\pi) + 1 + \ln\hat{\sigma}_u^2) + 2(m+1), \tag{3.112}$$

$$\mathrm{BIC} = T(\ln(2\pi) + 1 + \ln\hat{\sigma}_u^2) + (m+1)\ln(T). \tag{3.113}$$

We can select the regression with parameter estimates for the degree m as a model where the information criteria are minimised.

3.1.9.5 Multiple Linear Regression Analysis

Let y_t and $x_{i,t}$ be T sets of observations ($i = 1, \ldots, p$; $t = 1, \ldots, T$). Assuming that y_t is an explained variable and $x_{i,t}$ are explanatory variables, we consider their OLS regressions:

$$y_t = \beta_0 + \sum_{i=1}^{p}\beta_i x_{i,t} + z_t, \tag{3.114}$$

where $\beta_0, \ldots, \beta_p$ represent regression coefficients, and z_t an error. If we assume that the residual z_t is drawn from a zero-mean normal distribution with variance σ_z^2, $z_t \sim N\left(0, \sigma_z^2\right)$, then we can explicitly express the analytical solutions of regression

coefficients. Substituting observed data into the regression Eq. (3.114) and setting regression coefficients $\beta_0, \ldots, \beta_p$ to the values which minimise $\sum_{t=1}^{T} z_t^2$, we obtain the estimated parameters $\hat{\beta}_0, \ldots, \hat{\beta}_p$.

Let the density function of z be assumed as:

$$f(z; \sigma_z^2) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left[-\frac{z^2}{2\sigma_z^2}\right]. \tag{3.115}$$

Now, in order to estimate parameters $\beta_0, \ldots, \beta_p$, we attempt to maximise the log–likelihood value

$$L(\beta_0, \ldots, \beta_p, \sigma_z^2) = \sum_{t=1}^{T} \ln\Big(\frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left[-\frac{z_t^2}{2\sigma_z^2}\right]\Big), \tag{3.116}$$

with respect to $\beta_0, \ldots, \beta_p$ and σ_z^2. The parameters $\beta_0, \ldots, \beta_p$ and σ_z^2 are estimated as the maximum likelihood estimators

$$\{\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\sigma}_z^2\} = \arg \max_{\beta_0, \ldots, \beta_p, \sigma_z^2} L(\beta_0, \ldots, \beta_p, \sigma_z^2).$$

Partially differentiating L in terms of $\beta_0, \ldots, \beta_p$ and σ_z^2, and setting them into zero, i.e., $\frac{\delta L}{\delta \beta_i} = 0 \quad (i = 0, \ldots, p)$ and $\frac{\partial L}{\partial \sigma_z^2} = 0$, we get

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} T & \sum x_{1,t} & \sum x_{2,t} & \cdots & \sum x_{p,t} \\ \sum x_{1,t} & \sum x_{1,t}^2 & \sum x_{1,t}x_{2,t} & \cdots & \sum x_{1,t}x_{p,t} \\ \sum x_{2,t} & \sum x_{1,t}x_{2,t} & \sum x_{2,t}^2 & \cdots & \sum x_{2,t}x_{p,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_{p,t} & \sum x_{1,t}x_{p,t} & \sum x_{2,t}x_{p,t} & \cdots & \sum x_{p,t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum x_{1,t}y_t \\ \sum x_{2,t}y_t \\ \vdots \\ \sum x_{p,t}y_t \end{bmatrix}, \tag{3.117}$$

where $\sum$ stands for $\sum_{t=1}^{T}$, and

$$\hat{\sigma}_z^2(\hat{\beta}_0, \ldots, \hat{\beta}_p) = \frac{\sum_{t=1}^{T} z_t^2}{T} = \frac{\sum_{t=1}^{T}(y_t - \hat{\beta}_0 - \sum_{i=1}^{p} \hat{\beta}_i x_{i,t})^2}{T}. \tag{3.118}$$

Therefore, the maximised log–likelihood value turns out to be

$$\begin{aligned}\max L &= \max_{\beta_0,\ldots,\beta_p,\sigma_z^2} L(\beta_0,\ldots,\beta_p,\sigma_z^2)\\ &= L(\hat{\beta}_0,\ldots,\hat{\beta}_p,\hat{\sigma}_z^2)\\ &= -\frac{T}{2}\left(\ln(2\pi)+1+\ln\hat{\sigma}_z^2(\hat{\beta}_0,\ldots,\hat{\beta}_p)\right). \end{aligned} \tag{3.119}$$

Thus, AIC and BIC are written as

$$\mathrm{AIC} = T(\ln(2\pi)+1+\ln\hat{\sigma}_z^2(\hat{\beta}_0,\ldots,\hat{\beta}_p))+2(p+1), \tag{3.120}$$

$$\mathrm{BIC} = T(\ln(2\pi)+1+\ln\hat{\sigma}_z^2(\hat{\beta}_0,\ldots,\hat{\beta}_p))+(p+1)\ln(T). \tag{3.121}$$

3.1.10 Numerical Assessment of Sampling Error

For the probability distribution, statistics and simple aspects of the base stochastic variable, it may be possible to assess closely and analytically the error for a finite number of observations. However, this is generally difficult. While it is possible to approximate the sampling error by assuming a sufficiently large number of observations based on the central limit theorem, when the actual number of observations is small, it may not be possible to have confidence in this value.

The *bootstrap method* and the *jackknife method* are methods for reusing sampling data to assess the sampling error for this type of estimator [38]. These are vigorous methods that can be conducted regardless of the form of the probability distribution and complexity of the scope of estimation. The bootstrap method consists of approximately generating the sample distribution when the sample distribution of a statistic is unknown for an event that one wants to learn about to assess the sampling error and other factors of the statistic.

Assume that the sampling time series is $x_1,\ldots,x_T$. With the bootstrap method, the bootstrap sample $x_1^*,\ldots,x_T^*$ is composed of a time series allowed to be repeatedly randomly selected (sampling with replacement) from T items of sample data. Creating B bootstrap sequences, we compute a statistic computed from B statistics, denoted as θ_i^* $(i=1,\ldots,B)$. Through simulation, it is possible to estimate the sampling error from the sample mean and the sample standard deviation computed from

$$\theta^* = \frac{1}{B}\sum_{i=1}^{B}\theta_i^*, \tag{3.122}$$

$$\text{std. err.} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\theta_i^*-\theta^*)^2}. \tag{3.123}$$

With the jackknife method, proposed by Quenouille [35], a jackknife sample $x_1^*, \ldots, x_{T-T/m}^*$ is generated as follows. Let the ith group be divided into m groups of size T/m each. The statistic θ_i^* is computed from the jackknife sample where the ith group of size T/m has been deleted from the original sample time series $x_1, \ldots, x_T$.

The sampling error from the sample mean and sample standard deviation are computed from

$$\theta^* = \frac{1}{m}\sum_{i=1}^{m}\theta_i^*, \tag{3.124}$$

$$\text{std. err.} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\theta_i^* - \theta^*)^2}. \tag{3.125}$$

Generally, with the jackknife method the sampling error can be calculated with a smaller amount of calculation than the bootstrap method.

3.1.11 Statistical Hypothesis Testing

In some cases, it is necessary to make a statistical judgement using data to determine whether the hypothesis is correct. In this case, a formal procedure called a statistical hypothesis test is used.

In the statistical hypothesis test, first a hypothesis called a null hypothesis that is the opposite of what we wish to prove is established. The null hypothesis is written as H_0. In addition, a hypothesis that is in opposition to the null hypothesis is often established. This hypothesis in opposition to the null hypothesis is called an alternative hypothesis, which is written as H_1. Rejecting the null hypothesis H_0 indicates that the alternative hypothesis H_1 is true.

In general, the following procedure is used in the statistical hypothesis test:

1. Establish a null hypothesis and an alternative hypothesis.
2. Define a test statistic and seek the sample distribution.
3. Decide on a significance level and establish the rejection range for the test statistic.
4. Calculate the test statistic's instance from the sample data.
5. If the instance is in the rejection range, reject the null hypothesis and adopt the alternative hypothesis.

As shown in Table 3.2, regardless whether the null hypothesis H_0 is correct, there is the chance of the type I error of erroneously rejecting the null hypothesis and the type II error of erroneously not rejecting the null hypothesis.

It is impossible to decrease type I errors and type II errors at the same time. In response to this, the type I error is emphasised by selecting verification procedure that minimises the type II error under the restriction of not exceeding a significance

Table 3.2 Decision making in a statistical hypothesis test

	H_0 is true	H_1 is true
Null hypothesis H_0 is not rejected	Correct	Type II error
Null hypothesis H_0 is rejected	Type I error	Correct

level α (normally, α is set as 0.01, 0.05, etc.), that is determined in advance. This is called the Neyman–Pearson framework.

When a test has a significance level α, it indicates that the test formula will lower the probability of type I errors to below that level. In other words, the probability of erroneously rejecting the null hypothesis H_0 will fulfil $\Pr[H_0] \leq \alpha$ when it is correct.

Assume a sample $X = (X_1, \ldots, X_T)$, and use a statistic $Y = u(X_1, \ldots, X_T)$ to test the null hypothesis H_0. Assume that the rejection range for the test can be written as $u(x_1, \ldots, x_T) \leq c$ (in other words, reject H_0 if the rejection level c falls below the test statistic Y instance) and that the observation value is $y = u(x_1, \ldots, x_T)$. At this time, $p = \Pr[Y \leq y; H_0]$ is called p-value of the actual data. $\Pr[u(x_1, \ldots, x_T) \leq c; H_0] = 0.05$ means that H_0 is rejected by a significance level of 5 % if $y \leq c$.

If the rejection range is $u(x_1, ..., x_T) \geq c$, the p-value on the left tail of the distribution would be $p = \Pr[Y \geq y; H_0]$, or if the rejection range is $|u(x_1, \ldots, x_T)| \geq c$, it would result in $p = \Pr[|Y| \geq |y|; H_0]$ in both tails of the distribution. In this manner, a p-value viewed in one tail of the test distribution is called a single-sided p-value, and a p-value viewed in both tails is called a two-sided p-value.

In other words, because the actual value of the test statistic is $y = u(x_1, ..., x_T)$ if the data at hand $(x_1, \ldots, x_T)$ is used for calculation, if the stochastic variable $y = u(X_1, ..., X_T)$ occurs repeatedly over multiple trials when the null hypothesis is correct, this results in a probability assessment that exceeds result y, representing one trial. In this case, the p-value indicates just how extreme an instance the y-value is. Namely, the smaller the p-value is, the more it is probabilistically unlikely for y to occur, and if y does actually occur, there is a basis to doubt the truth of H_0.

3.1.12 Anderson–Darling and Kolmogorov–Smirnov Tests

In this subsection, statistical hypothesis tests to assess the difference between datasets and its assumed distribution are introduced.

Kolmogorov–Smirnov (KS) test is a popular statistical method of assessing a difference between datasets and its assumed distribution by p-value, which is a measure of probability where a difference between these two distributions happens by chance [39]. This test is widely used because it does not depend on the distribution form of data. However, the KS test has a disadvantage that it tends to be more sensitive near the centre of the distribution than at the tails.

Anderson–Darling (AD) test is an efficient method [5], which is a generalisation of the KS test. We can assess the statistical significance level putting a weight on the part of distributions which we think important. Here, we show how to assess the difference

between datasets and its assumed distribution focusing on tails, and calculate p-value. This will be used for evaluating parameter estimates for risk estimation in Sects. 5.6 and 5.7.

Suppose that one has T observations $\{x_1, x_2, \ldots, x_T\}$, and let K_T be a test statistic of the KS test or the AD test,

$$K_T = \sup_{1 \le i \le T} \sqrt{T} |P_T(x_i) - P(x_i)| \sqrt{\psi(P(x_i))} \ (0 \le P(x) \le 1), \tag{3.126}$$

where $P(x)$ is an assumed CDF, and $P_T(x)$ an empirical one based on T observations such that $P_T(x_i) = \frac{k}{T}$, in which k represents x_i's ascending order.

The test statistic K_T is the measure of distance between these two distributions, and we can put a weight on the deviations with $\psi(u)$ $(0 \le u \le 1)$ according to the importance attached to portions CDF. In the case of KS test, the weight function is chosen as $\psi(u) = 1$. p-value of the KS test is given as the Kolmogorov–Smirnov distribution:

$$p = \Pr[K \ge z] = 2 \sum_{n=1}^{\infty} (-1)^{n-1} e^{-2n^2 z^2}. \tag{3.127}$$

If $p < \alpha$ $(0 < \alpha < 1)$, then the hypothesis that an assumed probability distribution coincides the empirical probability distribution is rejected at level α.

In the case of AD test, $\psi(u) = \frac{1}{u(1-u)}$ is often assumed. Now we want to know the CDF of the test statistic K_T in order to obtain the p-value of the AD test. Consider the transformation $u = P(x)$. Under the null hypothesis that $\{x_i\}$ $(i = 1, \ldots, T)$ are drawn from the distribution $P(x)$, $\{u_i\}$ $(i = 1, \ldots, T)$ can be considered as drawn from the uniform distribution for $0 \le u \le 1$. Introducing $G_T(u)$ defined as the empirical distribution derived from $\{u_1, \ldots, u_T\}$, from Eq. (3.126) we get

$$K_T = \sup_{1 \le i \le T} \sqrt{T} |G_T(u_i) - u_i| \sqrt{\psi(u_i)}. \tag{3.128}$$

Then, $Y_T = \sqrt{T}(G_T(u) - u)$ is a random variable for $0 \le u \le 1$, and the set of these random variables may be considered to be a stochastic process with parameter u.

Let us assume

$$B_T(z) = \Pr\left\{ \sup_{1 \le i \le T} |Y_T(u_i)| \sqrt{\psi(u_i)} \le z \right\}, \tag{3.129}$$

and calculate $B(z) = \lim_{T \to \infty} B_T(z)$.
For $T \to \infty$, the limiting process of $Y_T(u)$ is a Gaussian process $y(u)$, $0 \le u \le 1$, which is specified as

$$\mathrm{E}[y(u)] = 0,$$
$$\mathrm{E}[y(u)y(v)] = \min(u, v) - uv.$$

When putting

$$b(z) = \Pr\left\{\sup_{0\le u\le 1} |y(u)|\sqrt{\psi(u)} \le z\right\}, \tag{3.130}$$

we obtain $b(z) = B(z)$, which leads to $b(z) = \Pr[K \le z]$ where $K = \sup_{0\le u\le 1} |y(u)|\sqrt{\psi(u)}$.

In order to calculate the limiting distribution, we consider the problem of a corresponding stochastic process. It is obvious that the event $\{K \le z\}$ is equivalent to the event $\left\{-z\psi(u)^{-\frac{1}{2}} \le y(u) \le z\psi(u)^{-\frac{1}{2}},\ 0 \le u \le 1\right\}$, which naturally leads to

$$\Pr\{K \le z\} = \Pr\left\{-z\psi(u)^{-\frac{1}{2}} \le y(u) \le z\psi(u)^{-\frac{1}{2}}\right\}. \tag{3.131}$$

Therefore, it can be very crudely said that the probability $b(z)$ is the proportion of all paths $y(u)$ of the diffusing particle which do not get absorbed into the barriers $y = \pm z\psi(u)^{-\frac{1}{2}}$.

It is convenient to make the following transformation which simplifies the analysis:

$$X(t) = (1 + t)y\Big(\frac{t}{1 + t}\Big), \tag{3.132}$$

where $X(t)$ is the Wiener-Einstein process for which

$$X(0) = 0, \quad \mathrm{E}[X(0)] = 0, \quad \mathrm{E}[X(s)X(t)] = \min(s, t).$$

Then, we can compute $\Pr[K \le z]$ from

$$\Pr[K \le z] = \Pr[|X(t)| \le \xi(t)\ ,\ 0 \le t \le \infty], \tag{3.133}$$

where

$$\xi(t) = \frac{z(1 + t)}{\sqrt{\psi(\frac{t}{1+t})}}.$$

Thus, we have the absorption probability problem for a free particle with barriers $X = \pm\xi(t)$ for $t \ge 0$.

One of the methods to solve the problem is to treat the corresponding diffusion problem as a boundary value problem with the Fokker–Planck equation for the diffusion process:

$$\frac{\partial f(t,X)}{\partial t} = \frac{1}{2}\frac{\partial^2 f(t,X)}{\partial X^2} \quad (t \geq 0,\ |X| \leq \xi(t)), \tag{3.134}$$

where $f(t, X)$ represents a time-dependent probability density function of X at time t and the initial condition $f(0, X) = \delta(X)$. In this analogy, $f(t', X)$ will be the density of $X(t')$ which have not been absorbed for $0 \leq t \leq t'$.
Then, we have

$$\Pr[K \leq z] = \lim_{t\to\infty} \int_{-\xi(t)}^{\xi(t)} f(t,X)\mathrm{d}X. \tag{3.135}$$

Eq. (3.127) is derived from Eq. (3.135) for $\psi(u) = 1$ [5].

To calculate Eq. (3.135) numerically, we consider the Crank-Nicholson discretisation of Eq. (3.134) such that

$$\begin{aligned}\frac{f(t_{l+1},X_j) - f(t_l,X_j)}{\Delta t} = \frac{1}{4(\Delta X)^2}\Big[\big(f(t_{l+1},X_{j+1}) - 2f(t_{l+1},X_j)\\ + f(t_{l+1},X_{j-1})\big) + \big(f(t_l,X_{j+1}) - 2f(t_l,X_j) + f(t_l,X_{j-1})\big)\Big],\end{aligned} \tag{3.136}$$

where $\Delta t = t_l - t_{l-1}$ and $\Delta X = X_j - X_{j-1}$ and $f(t_l, X_j)$ is obtained by the following tridiagonal problem:

$$\begin{aligned}-rf(t_{l+1},X_{j+1}) + (1+2r)f(t_{l+1},X_j) - rf(t_{l+1},X_{j-1}) =\\ rf(t_l,X_{j+1}) + (1-2r)f(t_l,X_j) + rf(t_l,X_{j-1}),\end{aligned} \tag{3.137}$$

where $r = \frac{\Delta t}{4(\Delta X)^2}$. Enumerating Eq. (3.137) iteratively for large enough l, with the boundary condition $f(t, X) = 0$ for $|X| \geq \xi(t)$, we get $b(z)$ approximately from the numerical computation of $\int_{|X|<\xi(t)} f(t,X)\mathrm{d}X$.

There is also another way to get $b(z)$ approximately, using Monte Carlo simulation: The corresponding stochastic differential equation is $\mathrm{d}X(t) = \mathrm{d}W$. Using the Euler–Maruyama scheme as its discretisation of the equation, we obtain

$$X(t_{l+1}) = X(t_l) + \sqrt{\Delta t}\Delta W, \tag{3.138}$$

where $\Delta t = t_{l+1} - t_l$ and ΔW is sampled from a zero mean standard normal distribution $N(0, 1)$.

In the same way, we can obtain the CDF of K_T with $\psi(u)$ in general. When we focus on tails of distributions, we set $\psi(u) = \frac{1}{u(1-u)}$. Then, the barriers $\xi(t)$ in the corresponding Wiener process is set as $\xi(t) = \sqrt{t}$. In the case of $\psi(u) = \frac{1}{u(1-u)}$, the range of t in the process is confined for numeration as $\frac{a}{1-a} < t < \frac{b}{1-b}$, where $a = P(\min_{1\leq i\leq T} x_i)$ and $b = P(\max_{1\leq i\leq T} x_i)$ [5].

Here, we introduce another convenient transformation to calculate $b(z)$:

$$X(t) = \sqrt{t}U\left(\frac{1}{2\alpha_X}\ln t\right). \tag{3.139}$$

Then, $U(t)$ is an Uhlenbeck process with correlation parameter α_X and

$$b(z) = \Pr\left\{|U(t)| \le z,\ 0 \le t \le \frac{1}{2\alpha_X}\ln\frac{b(1-a)}{a(1-b)}\right\}. \tag{3.140}$$

In similar to the Wiener process, the problem is solved as a boundary value problem with the corresponding Fokker–Planck equation,

$$\frac{\partial f'(t,U)}{\partial t} = \alpha_X\frac{\partial f'(t,U)}{\partial U} + \alpha_X\frac{\partial^2 f'(t,U)}{\partial U^2} \quad \left(|U| \le z,\ 0 \le t \le \frac{1}{2\alpha_X}\ln\frac{b(1-a)}{a(1-b)}\right), \tag{3.141}$$

where $f'(t,U)$ is the PDF that $U(t)$ have not been absorbed by the barriers, yet. Therefore, we describe the Crank–Nicholson discretisation of Eq. (3.141):

$$\begin{aligned}\frac{f'(t_{l+1},U_j) - f'(t_l,U_j)}{\Delta t} = \frac{\alpha_X}{2}\Big[&\frac{f'(t_{l+1},U_{j+1}) - f'(t_{l+1},U_{j-1})}{2(\Delta U)} \\ &+ \frac{f'(t_l,U_{j+1}) - f'(t_l,U_{j-1})}{2(\Delta U)} \\ &+ \frac{f'(t_{l+1},U_{j+1}) - 2f'(t_{l+1},U_j) + f'(t_{l+1},U_{j-1})}{(\Delta U)^2} \\ &+ \frac{f'(t_l,U_{j+1}) - 2f'(t_l,U_j) + f'(t_l,U_{j-1})}{(\Delta U)^2}\Big],\end{aligned} \tag{3.142}$$

and $b(z)$ is approximated as $\sum_j f'(t,U)\Delta U$ by using the solution $f'(t,U_j)$ computed from Eq. (3.142) at $t = \frac{1}{2\alpha_X}\ln\frac{b(1-a)}{a(1-b)}$ under the initial condition $f'(0,U) = \delta(U)$ and the boundary condition $f'(t,U) = 0$ for $|U| \ge z$. In the case of Monte Carlo simulation, the corresponding stochastic differential equation is $\mathrm{d}U(t) = -\alpha_X U(t)\mathrm{d}t + \sqrt{2\alpha_X}\mathrm{d}W$. Using the Euler–Maruyama scheme as the discretisation, we have

$$U_l = U_{l-1} - \alpha_X U_{l-1}\Delta t + \sqrt{2\alpha_X\Delta t}\,\Delta W, \tag{3.143}$$

and $b(z)$ is estimated as $\frac{n_a - n_b}{n_a}$, where n_b is the number of n_a samples that U_i is absorbed into the barrier for $0 \le t \le \frac{1}{2\alpha_X}\ln\frac{b(1-a)}{a(1-b)}$.

3.2 Time Series Analysis

In various fields, it is common to deal with data collected from sequential observations over time. In business, we observe exchange rates between pairwise currencies, stock prices, and interest rates. This is called a *time series*.

The number of events observed in a period is also represented as a time series. This is modelled as a *counting process*. For example, the daily number of available flights, the daily number of available hotels, and the number of transactions of currency exchange within each minute can be expressed as a time series, respectively.

This section will introduce fundamental methods to treat the time series are explained. It is a nontrivial task to find appropriate models of time series. A multi-step model-building strategy was proposed by Box and Jenkins [15]. There are three main steps, each of which may be used several times:

1. *Model specification* (or identification)
2. *Model fitting*
3. *Model diagnostics*

In the model specification step, we may select a model that may be appropriate for a given observed series. In this step, we often look at the plot of time series and compute many different statistics from the data.

To choose a model, we shall attempt to use the principle of parsimony. This is related to the following remarks by Albert Einstein in Parzen: 'Everything should be made as simple as possible but not simpler'. This is an assumption that the model used should require the smallest number of parameters that will adequately represent the time series. The information criterion is often used for this purpose.

In the model fitting step, the model parameters should be estimated from the data. We shall consider criteria such as least squares and maximum likelihood for estimation.

Model diagnostics is concerned with assessing the quality of the model that we have specified and estimated. We repeatedly compute statistics to find inadequacies between the data and the model. We cycle through the three steps until an acceptable model is found.

3.2.1 Stochastic Processes

Suppose a sequence of random variables $\{Y_t(\omega) : t = 0, \pm 1, \pm 2, \pm 3, \ldots\}$. This is called the *stochastic process*, and it serves as a model for an observed time series. ω represents unknown information or input, and it is ignored in many cases. Namely, let $\{Y_t : t = 0, \pm 1, \pm 2, \pm 3, \ldots\}$ be a stochastic process. The probabilistic structure of the stochastic process is given by the set of distributions of all finite collections of Y_t. If we accept the assumption that the information of joint probability distributions characterising the stochastic process can be described in terms of means, variances, and covariances, then our task is mainly to compute these first and second moments.

3.2.2 Means, Variances and Covariances

For a stochastic process $\{Y_t : t = 0, \pm 1, \pm 2, \pm 3, \ldots\}$, the mean is defined as

$$\mu_t = \mathrm{E}[Y_t], \tag{3.144}$$

where μ_t is the expected value of the process at time t. Generally, μ_t can differ at each time point t. The autocovariance function, $\gamma_{t,s}$, is defined as

$$\gamma_{t,s} = \mathrm{Cov}[Y_t, Y_s] \quad \text{for} \quad t, s = 0, \pm 1, \pm 2, \ldots, \tag{3.145}$$

where $\mathrm{Cov}[Y_t, Y_s] = \mathrm{E}[(Y_t - \mu_t)(Y_s - \mu_s)] = \mathrm{E}[Y_t Y_s] - \mu_t \mu_s$. Then, the autocorrelation function, $\rho_{t,s}$, is given by

$$\rho_{t,s} = \frac{\mathrm{Cov}[Y_t, Y_s]}{\sqrt{\mathrm{Var}[Y_t]\mathrm{Var}[Y_s]}} = \mathrm{Corr}[Y_t, Y_s] = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}. \tag{3.146}$$

The following important properties for $\gamma_{t,s}$ and $\rho_{t,s}$ are known:

$$\begin{cases} \gamma_{t,t} = \mathrm{Var}[Y_t] & \rho_{t,t} = 1 \\ \gamma_{t,s} = \gamma_{s,t} & \rho_{t,s} = \rho_{s,t} \\ |\gamma_{t,s}| \leq \sqrt{\gamma_{t,t}\gamma_{s,s}} & \rho_{t,s} \leq 1 \end{cases}. \tag{3.147}$$

Values of $\rho_{t,s}$ near ± 1 indicate strong (linear) dependence, while values near zero indicate weak (linear) dependence. If $\rho_{t,s} = 0$, it is said that Y_t and Y_s are uncorrelated.

To make statistical inferences about the structure of a stochastic process from the data, we must usually assume that the probability law governing the behaviour of the process does not change over time. This is called *stationarity*. This is equivalent to the assumption that the process is in statistical equilibrium. In this case, means and variances are constant over time: $\mathrm{E}[Y_t] = \mathrm{E}[Y_{t-k}]$ and $\mathrm{Var}[Y_t] = \mathrm{Var}[Y_{t-k}]$. Furthermore, the covariance between Y_t and Y_s can be described in terms of the time difference $|t - s|$:

$$\gamma_{t,s} = \gamma_{0,|t-s|}. \tag{3.148}$$

Therefore, for a stationary process, we use simple notations:

$$\gamma_k = \mathrm{Cov}[Y_t, Y_{t-k}], \quad \rho_k = \mathrm{Corr}[Y_t, Y_{t-k}]. \tag{3.149}$$

3.2.3 Autoregressive Model

An *autoregressive model* (AR model) is a linear model under the assumption that the current value of the series x_t can be represented as a linear combination of the p most recent past values of itself $\{x_{t-p}, \ldots, x_{t-1}\}$

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + u_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t, \tag{3.150}$$

where a_1, …, a_p are coefficients and u_t is an innovation term that is not explained by the past values. Thus, for every time t, u_t is independent of $x_{t-1}, x_{t-2}, x_{t-3}, \ldots$. Since an AR model is both simple and solvable, this has various applications in time series analysis.

Assuming that u_t are sampled from *i.i.d.* normal distributions with zero mean and variance σ_u^2, we can rewrite Eq. (3.150) as

$$u_t = -\sum_{i=0}^{p} a_i x_{t-i}, \tag{3.151}$$

where $a_0 = -1$ is assumed. Then, the log–likelihood function for $t \in [n_0, n_1]$ is described as

$$l(a_0, a_1, a_2, \ldots, a_p, \sigma_u^2) = -\frac{n_1 - n_0 + 1}{2} \ln(2\pi\sigma_u^2) - \frac{1}{2\sigma_u^2} \sum_{t=n_0}^{n_1} \left(\sum_{i=0}^{p} a_i x_{t-i} \right)^2. \tag{3.152}$$

Partially differentiating Eq. (3.152) in terms of parameters a_0, …, a_p and setting them at zero imply that

$$\frac{\partial l}{\partial a_0} = \frac{\partial l}{\partial a_1} = \frac{\partial l}{\partial a_2} = \cdots = \frac{\partial l}{\partial a_p} = 0. \tag{3.153}$$

Using $a_0 = -1$, we have

$$\begin{bmatrix} C(1,1) & C(1,2) & C(1,3) & \cdots & C(1,p) \\ C(2,1) & C(2,2) & C(2,3) & \cdots & C(2,p) \\ \vdots & \vdots & \ddots & & \vdots \\ C(p,1) & C(p,2) & C(p,3) & \cdots & C(p,p) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = - \begin{bmatrix} C(0,1) \\ C(0,2) \\ \vdots \\ C(0,p) \end{bmatrix}, \tag{3.154}$$

where

$$C(i,\, j) = \sum_{t=n_0}^{n_1} x_{t-i} x_{t-j}. \tag{3.155}$$

Equation (3.154) is called *Yule–Walker equation*. Furthermore, by solving $\frac{\partial l}{\partial \sigma_u^2} = 0$, we get the maximum likelihood estimator of σ_u^2 as follows:

$$\hat{\sigma}_u^2 = \frac{1}{n_1 - n_0 + 1} \sum_{t=n_0}^{n_1} \Big(\sum_{i=0}^{p} a_i x_{t-i} \Big)^2 = \frac{1}{n_1 - n_0 + 1} \sum_{i=0}^{p} \sum_{j=0}^{p} a_i a_j C(i,\, j). \tag{3.156}$$

Thus, the maximum log–likelihood value is given as

$$l(\hat{a}_0, \ldots, \hat{a}_p, \sigma_u^2) = -\frac{1}{2}(n_1 - n_0 + 1)(\ln(2\pi) + 1 + \ln \hat{\sigma}_u^2). \tag{3.157}$$

Therefore, information criteria (AIC and BIC) can be derived as

$$\mathrm{AIC} = (n_1 - n_0 + 1)(\ln(2\pi) + 1 + \ln \hat{\sigma}_u^2) + 2(p + 1), \tag{3.158}$$

$$\mathrm{BIC} = (n_1 - n_0 + 1)(\ln(2\pi) + 1 + \ln \hat{\sigma}_u^2) + (p + 1) \ln(n_1 - n_0 + 1). \tag{3.159}$$

If the number of observations is T, then n_1 is set as T. Since x_t is determined by the most recent past values of itself $\{x_{t-p}, \ldots, x_{t-1}\}$, $n_0 - 1$ should be greater than or equal to p.

3.2.4 Segmented Regression Analysis

We consider how we can improve the precision of multiple linear regression analysis mentioned in Sect. 3.1.9.5. Normally, we use all the data for the given period and conduct a linear regression analysis. However, this does not work well when the trend of the data changes during the period. If we can find the turning points, we should divide the data into a few segments at those points. We can improve the precision by conducting linear regression analysis for each segment of data sets because it corresponds to each tendency. Furthermore, when we want to predict the future realisation even in a short period, we may use the latest data set because it includes the latest trend. Such a method is called *segmented regression* [25, 44]. We discuss how we can find the turning points of data in terms of tendency, and show the segmented linear regression analysis below.

Let τ $(1 < \tau < T)$ be the point to divide the data into two segments, that is, $t_l \in \{1, \ldots, \tau\}$, $t_r \in \{\tau + 1, \ldots, T\}$. Here, in order to regard heteroscedasticity of disturbances z_t, we define an alternative log–likelihood value:

$$L_2(\tau; \beta_{l0}, \ldots, \beta_{lp}, \sigma_l^2, \beta_{r0}, \ldots, \beta_{rp}, \sigma_r^2)$$
$$= \sum_{t=1}^{\tau} \ln \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left[-\frac{z_{lt}^2}{2\sigma_l^2}\right] + \sum_{t=\tau+1}^{T} \ln \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left[-\frac{z_{rt}^2}{2\sigma_r^2}\right], \tag{3.160}$$

where

$$z_{lt} = y_t - \beta_{l0} - \sum_{i=1}^{p} \beta_{li} x_{i,t}, \tag{3.161}$$

$$z_{rt} = y_t - \beta_{r0} - \sum_{i=1}^{p} \beta_{ri} x_{i,t}. \tag{3.162}$$

The parameters $\beta_{l0}, \ldots, \beta_{lp}, \sigma_l^2, \beta_{r0}, \ldots, \beta_{rp}, \sigma_r^2$ are estimated as the maximum likelihood estimators

$$\{\hat{\beta}_{l0}, \ldots, \hat{\beta}_{lp}, \hat{\sigma}_l^2, \hat{\beta}_{r0}, \ldots, \hat{\beta}_{rp}, \hat{\sigma}_r^2\}$$
$$= \arg \max_{\beta_{l0},\ldots,\beta_{lp},\sigma_l^2,\beta_{r0},\ldots,\beta_{rp},\sigma_r^2} L_2(\tau; \beta_{l0}, \ldots, \beta_{lp}, \sigma_l^2, \beta_{r0}, \ldots, \beta_{rp}, \sigma_r^2).$$

Partially differentiating $L_2(\tau; \beta_{l0}, \ldots, \beta_{lp}, \sigma_l^2, \beta_{r0}, \ldots, \beta_{rp}, \sigma_r^2)$ in terms of its parameters conditioning on fixed τ and setting them into zero, we have

$$\begin{bmatrix} \hat{\beta}_{l0} \\ \hat{\beta}_{l1} \\ \vdots \\ \hat{\beta}_{lp} \end{bmatrix} = \begin{bmatrix} \tau & \sum x_{1,t} & \sum x_{2,t} & \cdots & \sum x_{p,t} \\ \sum x_{1,t} & \sum x_{1,t}^2 & \sum x_{1,t}x_{2,t} & \cdots & \sum x_{1,t}x_{p,t} \\ \sum x_{2,t} & \sum x_{1,t}x_{2,t} & \sum x_{2,t}^2 & \cdots & \sum x_{2,t}x_{p,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{p,t} & \sum x_{1,t}x_{p,t} & \sum x_{2,t}x_{p,t} & \cdots & \sum x_{p,t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum x_{1,t}y_t \\ \sum x_{2,t}y_t \\ \vdots \\ \sum x_{p,t}y_t \end{bmatrix},$$

where $\sum$ stands for $\sum_{t=1}^{\tau}$, and

$$\hat{\sigma}_l^2(\hat{\beta_{l0}}, \cdots, \hat{\beta_{lp}}) = \frac{\sum_{t=1}^{\tau} \left(y_t - \hat{\beta}_{l0} - \sum_{i=1}^{p} \hat{\beta}_{li} x_{i,t}\right)^2}{\tau},$$

$$\begin{bmatrix} \hat{\beta}_{r0} \\ \hat{\beta}_{r1} \\ \vdots \\ \hat{\beta}_{rp} \end{bmatrix} = \begin{bmatrix} T-\tau & \sum x_{1,t} & \sum x_{2,t} & \cdots & \sum x_{p,t} \\ \sum x_{1,t} & \sum x_{1,t}^2 & \sum x_{1,t}x_{2,t} & \cdots & \sum x_{1,t}x_{p,t} \\ \sum x_{2,t} & \sum x_{1,t}x_{2,t} & \sum x_{2,t}^2 & \cdots & \sum x_{2,t}x_{p,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{p,t} & \sum x_{1,t}x_{p,t} & \sum x_{2,t}x_{p,t} & \cdots & \sum x_{p,t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum x_{1,t}y_t \\ \sum x_{2,t}y_t \\ \vdots \\ \sum x_{p,t}y_t \end{bmatrix},$$

where $\sum$ stands for $\sum_{t=\tau+1}^{T}$, and

$$\hat{\sigma}_r^2(\hat{\beta_{r0}}, \cdots, \hat{\beta_{rp}}) = \frac{\sum_{t=\tau+1}^{T} \left(y_t - \hat{\beta}_{r0} - \sum_{i=1}^{p} \hat{\beta}_{ri} x_{i,t}\right)^2}{T - \tau}.$$

Therefore, the maximum log–likelihood value is given as

$$\begin{aligned}
&\max L_2(\tau) \\
&= \max_{\beta_{l0},\ldots,\beta_{lp},\sigma_l^2,\beta_{r0},\ldots,\beta_{rp},\sigma_r^2} L_2(\tau; \beta_{l0}, \ldots, \beta_{lp}, \sigma_l^2, \beta_{r0}, \ldots, \beta_{rp}, \sigma_r^2) \\
&= L_2(\tau; \hat{\beta}_{l0}, \ldots, \hat{\beta}_{lp}, \hat{\sigma}_l^2, \hat{\beta}_{r0}, \ldots, \hat{\beta}_{rp}, \hat{\sigma}_r^2) \\
&= -\frac{T}{2}\big(\ln(2\pi) + 1\big) - \frac{\tau}{2} \ln \hat{\sigma}_l^2(\hat{\beta}_{l0}, \ldots, \hat{\beta}_{lp}) \\
&\quad -\frac{T - \tau}{2} \ln \hat{\sigma}_r^2(\hat{\beta}_{r0}, \ldots, \hat{\beta}_{rp}).
\end{aligned} \tag{3.163}$$

We, respectively, assume a null model and an alternative model as

H_0: OLS regression for all the data given in Eq. (3.114) (homogeneous disturbance distribution)
H_1: OLS regression for two segmented data at τ (a mixture of two different normal distributions)

Which OLS method is more suitable, the null model H_0, or the alternative model H_1? In this case, a likelihood-ratio test [46] may provide an answer to this question. Namely, the difference of log–likelihood can be used as a discriminant measure.

Let us introduce a logarithmic form of a likelihood-ratio $\Delta(\tau)$, the difference of maximum log–likelihood value between $L_2(\tau)$ and L, which are, respectively, defined in Eqs. (3.163) and (3.119),

$$\begin{aligned}
\Delta(\tau) &= \max L_2(\tau) - \max L \\
&= \frac{1}{2}\Big(T \ln \hat{\sigma}_z^2(\hat{\beta}_0, \ldots, \hat{\beta}_p) - \tau \ln \hat{\sigma}_l^2(\hat{\beta}_{l0}, \ldots, \hat{\beta}_{lp}) \\
&\quad -(T - \tau) \ln \hat{\sigma}_r^2(\hat{\beta}_{r0}, \ldots, \hat{\beta}_{rp})\Big).
\end{aligned} \tag{3.164}$$

More precisely, an information criterion is used as the discriminant measure. The difference of AIC between the null model and the alternative model is written as

$$\begin{aligned}
\Delta_{AIC}(\tau) &= -2 \max L_2(\tau) + 2 \times 2(p + 1) - (-2 \max L + 2 \times (p + 1)) \\
&= -2\Delta(\tau) + 2(p + 1).
\end{aligned} \tag{3.165}$$

We move tentative segmenting point τ and calculate for each data set the maximum log–likelihood value max L, which is described as above. In order to find the turning

point regarding tendency, we deal with the variable τ ranging from 5 to $T-4$. From this calculation, we obtain spectrum of $\Delta(\tau)$. According to the likelihood-ratio test, tentative segmenting point τ which maximises the log–likelihood ratio $\Delta(\tau)$ can be regarded as the most probable turning point:

$$\tau^* = \arg\max_{\tau} \Delta(\tau). \tag{3.166}$$

This is equivalent to

$$\tau^* = \arg\min_{\tau} \Delta_{AIC}(\tau). \tag{3.167}$$

Normally, a statistic computed from a finite number of data points is noise-dressed. Therefore, when we get τ^*, we must also compute a confidence level of τ^*. From the Wilks's theorem [46], it is known that an asymptotic distribution of $2\Delta(\tau)$ is χ^2 distribution. However, the distribution of $2\Delta(\tau)$ is empirically estimated by using the bootstrap method or jackknife method for small T, empirically. Let us discuss the confidential level of dividing the data when the threshold of the likelihood-ratio test is set to Δ_{th}. When $\Delta(\tau^*) \geq \Delta_{th}$, τ^* seems to be the segment boundary but not always. This is because τ^* contains a noise. Therefore, we should compute a confidence level of τ^* while considering how τ^* fluctuates. Such a statistical fluctuation can be approximated with a *Bootstrap method* [18].

Generally, the procedure for obtaining a *Bootstrap distribution* for the one-sample problem is as follows. When the data sample $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ is given, we construct the sample probability distribution $\hat{F}$. With $\hat{F}$ fixed, we draw a random sample of size T from $\hat{F}$, say

$$X_i^* = x_j, X_i^* \sim_{ind} \hat{F}, \tag{3.168}$$

where $i = 1, 2, \ldots, T$.

we call this the *Bootstrap sample*, $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_T^*)$ resampled from $\mathbf{x} = (x_1, x_2, \ldots, x_T)$. Then we can approximate the sampling distribution of $R(\mathbf{X}, F)$ by the *Bootstrap distribution* of

$$R^* = R(\mathbf{X}^*, \hat{F}), \tag{3.169}$$

i.e., the distribution of R^* is induced by the random mechanism Eq. (3.168) with $\hat{F}$ held fixed at its observed value. The simplest way to set probabilities to resample sequences is to put the probability $1/T$ at each point of $\boldsymbol{x}$.

Let $\mathbf{D}$ be the $T \times (p+1)$ data matrix and D_k represent the k-th data set, i.e., $\begin{bmatrix} x_{k,1} \ldots x_{k,p} & y_k \end{bmatrix}^T$. In our model, we conduct the *Bootstrap sequences* by the following procedure.

As we express in Fig. 3.4, we select τ^* data sets at random with replacement from the data sets $\mathbf{D}$ and call the obtained data sets $\mathbf{D}_l^*$. We introduce its t-th data set D_{lt}^* so that $D_{lt}^* = D_k \quad (t = 1, \ldots, \tau^*)$ where $k \in \{1, \ldots, \tau^*\}$ is selected with the probability $1/\tau^*$. We also select $(T - \tau^*)$ data sets from $\mathbf{D}$ and call it $\mathbf{D}_r^*$. Its

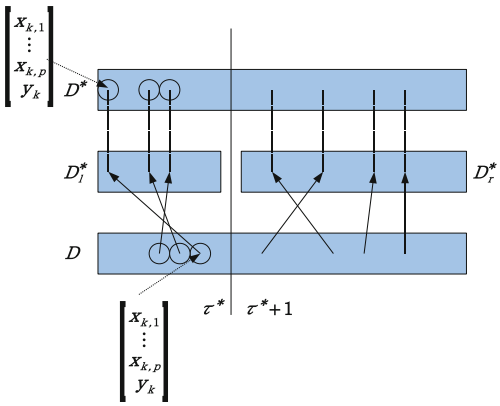


Fig. 3.4 Procedure to construct *Bootstrap sequences* $\mathbf{D}^*$, $\mathbf{D}_l^*$ and $\mathbf{D}_r^*$ from $\mathbf{D}$

t-th content will be $D_{rt}^* = D_k \quad (t = \tau^* + 1, \ldots, T)$ where $k \in \{\tau^* + 1, \ldots, T\}$ is selected with the probability $1/(T - \tau^*)$. We connect the obtained data sets as $\mathbf{D}^* = \left[\mathbf{D}_l^* \mathbf{D}_r^*\right]$.

Then, we calculate the disturbance terms z_t^*, z_{lt}^* and z_{rt}^* which we get by substituting data sets $\mathbf{D}^*$, $\mathbf{D}_l^*$ and $\mathbf{D}_r^*$ for Eqs. (3.114), (3.161) and (3.162). Using z_t^*, z_{lt}^* and z_{rt}^*, we can calculate max L^* and max $L_2^*(\tau^*)$ by Eqs. (3.119) and (3.163). At the end, we get the *Bootstrap variable* $\Delta^*(\tau^*)$ by substituting max L^* and max $L_2^*(\tau^*)$ for Eq. (3.164). Repeating this procedure, we can estimate the sampling distribution of $\Delta(\tau^*)$ by means of the *Bootstrap distribution*.

Figure 3.5 is an example of a *Bootstrap distribution*. As shown in this figure, we will be able to estimate the range of the error of $\Delta(\tau^*)$ with this distribution. We can estimate the value $\Delta_{5\ \%}$ at which its CDF is 5 %, as is described with a green line in Fig. 3.5. Thus we can set the condition to terminate segmenting recursively by using $\Delta_{5\ \%}$. Throughout this analysis, we set a threshold of segmenting as zero ($\Delta_{th} = 0$), and if $\Delta_{th} < \Delta_{5\ \%}$, we regard the segmenting procedure significant and continue it recursively, or otherwise, terminate segmenting for the considering interval. We use this for the terminal condition of segmentation.

Finally, following to this procedure, we get s segments and regression coefficients for each segments. We should evaluate if this model is well-fitted to the data given or not. Akaike information criterion (AIC) will give an answer to this. According to AIC, we should choose a model which is well-fitted to data with a small number of parameters as an adequate model from many candidates of models. AIC is defined as $AIC = -2 \times$ (log–likelihood) $+ 2 \times$ (the number of parameters). In the case of OLS, when the number of explanatory variables is p, AIC is given as:

$$AIC_{OLS} = -2 \times \max L + 2 \times (p + 1). \tag{3.170}$$

In the case of our proposed method, AIC is given as:

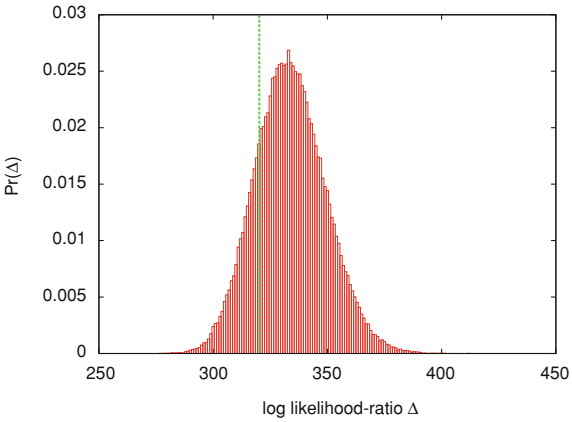


Fig. 3.5 *Bootstrap distribution* and its 5 % point. The *red boxes* represent the percentage of taking the value of log likelihood-ratio. The *green line* represents the value where the cumulative percentage is 5 %

$$AIC_{SOLS} = -2 \times \sum_{m=1}^{s} \max L_m + 2 \times \Big(s(p+1) + (s-1)\Big) \tag{3.171}$$

$$= -2 \sum_{m=1}^{s} \max L_m + 2\Big(s(p+2) - 1\Big), \tag{3.172}$$

where p is the number of explanatory variables, s is the number of segments, and $\max L_m$ is the maximum log–likelihood value computed from the m-th segmented period. We need "the number of segmented period $-$ 1" term because not only the number of explanatory variables but also the segmenting points τ^* are parameters.

3.3 Network Analysis

In this section, we will address fundamental parts of graph theory. They are needed when we attempt to express and analyse relationship of socioeconomic-technological systems. The network perspective provides new tools for answering questions in standard social and behavioural science research by giving precise formal definition to fields of social sciences. In addition, mobility of human behaviour and transportation are also represented using network description. Readers may find useful information in a number of review articles and books on complex networks [2, 8, 12, 43].

3.3.1 Basic Graph Theory

Graph theory has been commonly used in social network analysis to express the structure of relationships simply. Fundamentally, a graph consists of *nodes* (e) and of *links* (l) that connect the nodes. In the context of social sciences, nodes correspond

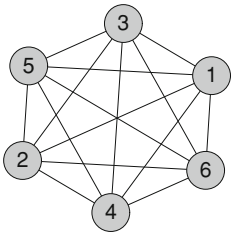


Fig. 3.6 The example of a fully connected network consisting of six nodes

to actors or events, and links represent the set of their relationships. We assume a set of nodes $E = \{e_1, \ldots, e_N\}$ and a set of links $V = (l_1, \ldots, l_L)$, where N and L are the number of nodes and the number of links, respectively.

The link is defined as the connection between two nodes. If a connection l_1 refers to the connection between nodes e_2 and e_6, we write as

$$l_1 = \{e_2, e_6\}. \tag{3.173}$$

For a network with the number of nodes equal to N, the maximum number of connections in an undirectional graph is given as

$$L_{max,undir} = \frac{N(N-1)}{2}. \tag{3.174}$$

The ratio of the total number of links to the maximum number of links is called *density* , which is defined as

$$d = \frac{2L}{N(N-1)}. \tag{3.175}$$

Figure 3.6 shows fully connected network consisting of six nodes. In this case, from Eq. (3.174), the number of links is calculated as $L = 6 \times 5/2 = 15$. This provides the maximum number of links L_{max} allowed in an undirectional graph.

Up to this point, a connection between two nodes is established, and the relationship is not in any specific directions. A directional connection is that which represents a connection that goes from one node (origin) and ends at another (destination). For example, if we try to conduct an analysis of a transportation network, there is a direction in the connections. Thus, if a connection l_1 refers to the directional connection of nodes e_2 and e_5,

$$l_1 = \{e_2 \rightarrow e_5\}. \tag{3.176}$$

In the case of directional connections, the maximum number of links is given as

$$L_{max,dir} = N(N-1). \tag{3.177}$$

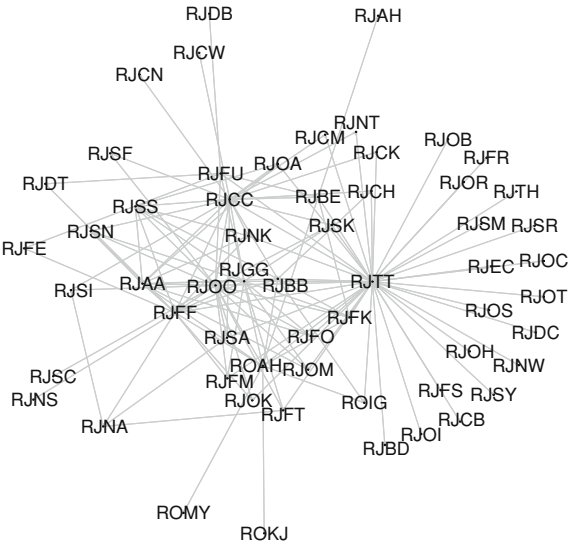


Fig. 3.7 The domestic air transportation network of Japan. This network consists of 58 nodes and is drawn from data for Japanese domestic flights on 15 October 2013 (JST). A node represents an airport, and a link a flight between two airports. 4-letter codes represent ICAO airport identifiers shown in Sect. 4.5 (p. 169)

Therefore, the density is calculated as

$$d = \frac{L}{N(N-1)}. \tag{3.178}$$

As an example of a directional graph in a real world, let us consider a Japanese domestic air transportation network. Figure 3.7 shows an example of the air transportation network of Japan. In this network, there are 58 airports ($N = 58$), represented as nodes that are connected with 261 links ($L = 261$). Since the maximum number of possible links is computed as $L_{max,dir} = 58 \times (58 - 1) = 3{,}306$, the density is estimated to be $d = 261/3{,}306 \times 100 = 7.89\,\%$. This implies that 7.89 % of the possible connections between two airports are used in Japanese domestic air transportation.

Graphs enable many interesting analyses to be made, and they have visual appeal, which helps us to understand the network. However, there are too many nodes and links to show in a visual representation. When increasing the number of nodes and links, it becomes impossible to use visualisation. Moreover, some important information, such as the frequency of occurrence and weights to characterise some information, are difficult to show in a graph. To solve this problem, we use the matrices to express the existence of a link between two nodes. Both directed and undirected networks can be expressed as a matrix. If the weights of links are homogeneous, then such a matrix is called an *adjacency matrix*.

Let A_{ij} represent an adjacency matrix. The indices i and j correspond to nodes e_i and e_j included in the set of nodes E. If the nodes e_i and e_j are connected, then an element of the matrix A_{ij} is unity. Otherwise, A_{ij} equals zero. For an undirectional

network, we have a symmetric matrix

$$A_{ij} = \begin{cases} 0 \ (e_i \text{ and } e_j \text{ are not connected}) \\ 1 \ (e_i \text{ and } e_j \text{ are connected}) \end{cases}, \tag{3.179}$$

and $A_{ij} = A_{ji}$. The diagonal element $A_{ii} = 0$ if we do not regard a self-connection of the node e_i. Figure 3.8a shows an example of a directed network for which an adjacency matrix is described as

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \tag{3.180}$$

We consider a *degree* of the i-th node, which is defined as

$$k_i = \sum_{j=1}^{N} A_{ij}. \tag{3.181}$$

The *degree sum formula* states that

$$\sum_{i=1}^{N} k_i = 2L. \tag{3.182}$$

We also define the *average degree* as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N}. \tag{3.183}$$

The average degree is sometimes called the density since there is a relationship between the average degree and the density defined in Eq. (3.175):

$$\langle k \rangle = d(N-1). \tag{3.184}$$

In the case of an undirected network, which is represented as Eq. (3.180), the degrees are given as $(k_1, k_2, k_3, k_4) = (2, 2, 1, 3)$. We can confirm that the sum of degrees is satisfied as

$$\sum_{i=1}^{4} k_i = 2 + 2 + 1 + 3 = 8 = 2L, \tag{3.185}$$

and compute the average degree as $\langle k \rangle = 8/4 = 2$ and the density as $d = 2/3$.

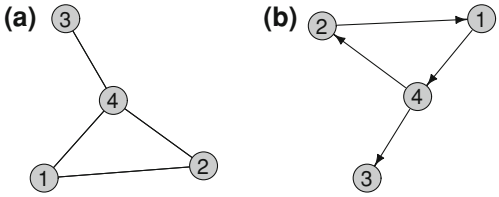


Fig. 3.8 An example of an undirected network (**a**) and a directed network (**b**)

If the connections are directional, we have an asymmetric matrix:

$$A_{ij} = \begin{cases} 0 \ (\{e_i \to e_j\} \text{ are not connected}) \\ 1 \ (\{e_i \to e_j\} \text{ are connected}) \end{cases}, \tag{3.186}$$

Figure 3.8b shows an example of directed networks and their adjacency matrix, for which an adjacency matrix is described as

$$\boldsymbol{A} = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,1\,1\,0 \end{pmatrix}. \tag{3.187}$$

In this case, we need to consider two types of degrees, called *in-degree* and *out-degree*. The in-degree is the number of links that enter a node. The out-degree is the number of links that emerge from a node. The in-degree and out-degree of the i-th node are, respectively, defined as

$$k_j^{(in)} = \sum_{i=1}^{N} A_{ij}, \tag{3.188}$$

$$k_i^{(out)} = \sum_{j=1}^{N} A_{ij}. \tag{3.189}$$

The average in-degree and the average out-degree are defined as

$$\langle k^{(in)} \rangle = \frac{1}{N} \sum_{j=1}^{N} k_j^{(in)}, \tag{3.190}$$

$$\langle k^{(out)} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{(out)}. \tag{3.191}$$

In the directed network, we have the degree sum formula for both in-degree and out-degree, such that

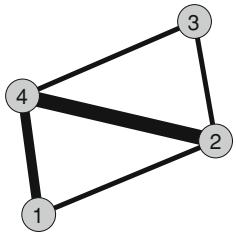


Fig. 3.9 An example of a weighted network consisting of four nodes

$$\sum_{j=1}^{N} k_j^{(in)} = \sum_{i=1}^{N} k_i^{(out)} = L. \tag{3.192}$$

Therefore, the average out-degree equals the average in-degree, which is equal to

$$\langle k^{(in)} \rangle = \langle k^{(out)} \rangle = \frac{L}{N} = d(N-1). \tag{3.193}$$

Similar to Eq. (3.184), we also find a relationship between the average degree and the density in the case of directed network. In the case of the directed network given in Eq. (3.187), the in-degrees and out-degrees are, respectively, computed as $(k_1^{(in)}, k_2^{(in)}, k_3^{(in)}, k_4^{(in)}) = (1, 1, 1, 1)$ and $(k_1^{(out)}, k_2^{(out)}, k_3^{(out)}, k_4^{(out)}) = (1, 1, 0, 2)$. The degree sum formula in the directed network is confirmed from the calculation:

$$\sum_{j=1}^{4} k_j^{(in)} = 1 + 1 + 1 + 1 = \sum_{i=1}^{4} k_i^{(out)} = 1 + 1 + 0 + 2 = 4 = L. \tag{3.194}$$

Therefore, the average in-degree, the average out-degree and the density are calculated as $\langle k^{(in)} \rangle = 1$, $\langle k^{(out)} \rangle = 1$, and $d = 1/3$.

The adjacency matrix expresses only connections. We often use a weighted adjacency matrix. One simple example of the weighted network is a matrix to express a flow (the yearly number of meetings between two persons, the daily number of buses travelling between two bus stops, the monthly number of flights between airports, and so forth) between e_i and e_j. Figure 3.9 shows an example of an undirected weighted network, for which a weighted adjacency matrix is described as

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 \\ 2 & 3 & 1 & 0 \end{pmatrix}. \tag{3.195}$$

This undirected graph consists of four nodes. For example, the weight of $l(1, 4)$ is given as 2 and $l(2, 3)$ as 1.

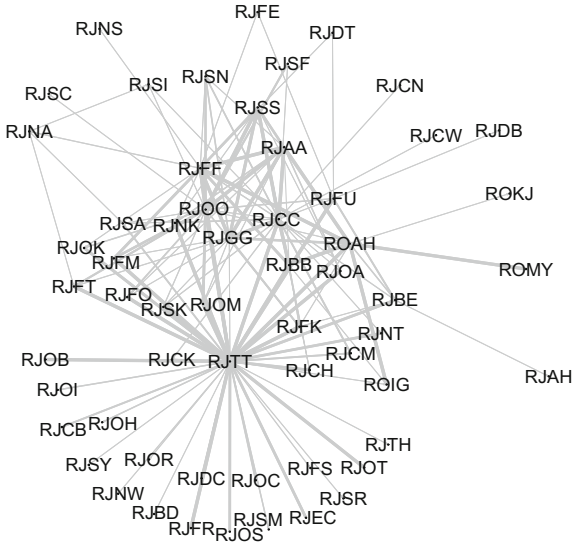


Fig. 3.10 The domestic air transportation network of Japan. This network consists of 58 nodes and is drawn from data for Japanese domestic flights on 15 October 2013 (JST). A node represents an airport, and a link a flight between two airports. The thickness of the link indicates the daily number of flights between departure and arrival airports. 4-letters codes indicate ICAO airport identifiers shown in Sect. 4.5 (p. 169)

As an example of a directed weighted network obtained from actual network data, we consider the Japanese domestic air transportation network shown in Fig. 3.10 again. The daily number of connections seems to be associated with the weight of a link. If two aeroplanes fly from Narita International Airport (RJAA) to Sendi Airport (RJSS) a day, we assign the weight from RJAA to RJSS as 2. Figure 3.7 shows the Japanese domestic air transportation network. The link weight is drawn in proportion to the daily number of connections between departure and arrival airports. We can see that there are many links among major airports, such as Narita International Airport (RJAA), Kansai International Airport (RJBB), New Chitose Airport (RJCC), Fukuoka Airport (RJFF), Osaka Airport (RJOO), Haneda Airport (RJTT) and so on.

3.3.2 Bipartite Graph

A *bipartite network* is a graph consisting of two types of nodes. Different types of nodes have links, but the same types of nodes have no links. Assume that there are two types of nodes, a and b. There are N nodes belonging to type a and M nodes belonging to type b. This structure can be expressed as an $(N + M) \times (N + M)$ adjacency matrix A_{ij}.

Suppose an undirectional bipartite network described by the adjacency matrix

$$\boldsymbol{A} = \begin{matrix} \\ N \\ \\ M \\ \\ \end{matrix} \overset{\overbrace{\quad}^{N}\ \overbrace{\quad}^{M}}{\begin{pmatrix} 0 \cdots 0 & 1 \cdots 0 \\ \vdots \ddots \vdots & \vdots \ddots \vdots \\ 0 \cdots 0 & 1 \cdots 0 \\ 1 \cdots 1 & 0 \cdots 0 \\ \vdots \ddots \vdots & \vdots \ddots \vdots \\ 0 \cdots 1 & 0 \cdots 0 \end{pmatrix}} \tag{3.196}$$

The first N nodes are included in type a, and the last M nodes are included in type b. The adjacency matrix A_{ij} can be partitioned into four blocks:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O}(N, N) & \boldsymbol{B} \\ \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{O}(M, M) \end{bmatrix}, \tag{3.197}$$

where the $N \times N$ block matrix and the $M \times M$ block matrix are zero matrix since there are no links between nodes of the same type. The matrix $\boldsymbol{B}$ is an $N \times M$ rectangular matrix in general. In the rectangular matrix $\boldsymbol{B}$, each column expresses a and each row b, and the element is 1 if the node belonging to b in that row has a link to the node belonging to a.

A degree of the i-th node belonging to a is given by $k_i^{(a)} = \sum_{j=1}^{M} B_{ij}$ $(i = 1, \ldots N)$, and a degree of the j-th node belonging to b is given by $k_j^{(b)} = \sum_{i=1}^{N} B_{ij}$ $(j = 1, \ldots, M)$.

In general, we can find the overlap for any pair of nodes a or b by summing the multiplied elements of the corresponding rows or columns of the rectangular adjacency matrix B_{ij}. That is,

$$A_{ij}^{(a)} = \sum_{k=1}^{M} B_{ik} B_{jk}, \quad A_{ij}^{(b)} = \sum_{k=1}^{N} B_{ki} B_{kj}, \tag{3.198}$$

where the matrices $\boldsymbol{A}^{(a)}$ and $\boldsymbol{A}^{(b)}$ give a one-mode projection by nodes belonging to type a and to type b, respectively. In matrix notation, Eq. (3.198) is rewritten as

$$\boldsymbol{A}^{(a)} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}, \quad \boldsymbol{A}^{(b)} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}. \tag{3.199}$$

The diagonal components of these matrices show the number of links between two nodes. For example, suppose 3 nodes of type a $(N = 3)$ and 4 nodes of type b $(M = 4)$, which construct a bipartite network described as the rectangular adjacency matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{3.200}$$

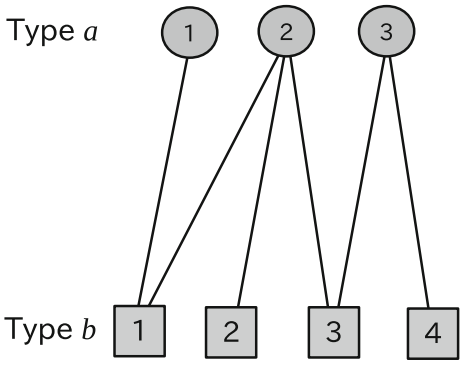


Fig. 3.11 An example of a bipartite graph

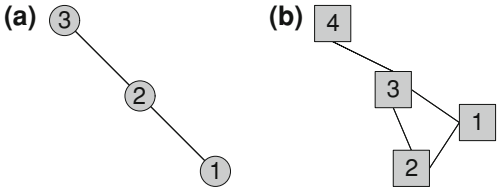


Fig. 3.12 One-mode projections of bipartite networks. **a** Projections by the nodes belonging to type a and **b** to type b

This bipartite network can be described as Fig. 3.11. The one-mode projection by nodes belonging to type a and to type b are given as

$$\boldsymbol{A}^{(a)} = \begin{bmatrix} 1\,0\,0\,0 \\ 1\,1\,1\,0 \\ 0\,0\,1\,1 \end{bmatrix} \begin{bmatrix} 1\,1\,0 \\ 0\,1\,0 \\ 0\,1\,1 \\ 0\,0\,1 \end{bmatrix} = \begin{bmatrix} 1\,1\,0 \\ 1\,3\,1 \\ 0\,1\,2 \end{bmatrix}, \tag{3.201}$$

and

$$\boldsymbol{A}^{(b)} = \begin{bmatrix} 1\,1\,0 \\ 0\,1\,0 \\ 0\,1\,1 \\ 0\,0\,1 \end{bmatrix} \begin{bmatrix} 1\,0\,0\,0 \\ 1\,1\,1\,0 \\ 0\,0\,1\,1 \end{bmatrix} = \begin{bmatrix} 2\,1\,1\,0 \\ 1\,1\,1\,0 \\ 1\,1\,2\,1 \\ 0\,0\,1\,1 \end{bmatrix}. \tag{3.202}$$

Non-diagonal elements of $A_{ij}^{(a)}$ and $A_{ij}^{(b)}$ show connections of one-mode projections by the nodes belonging to type a (see Fig. 3.12a) and to type b (see Fig. 3.12b), respectively. Figure 3.12 shows one-mode projection networks expressed as adjacency matrices, Eqs. (3.201) and (3.202). These are obtained as one-mode projections from a bipartite network shown in Fig. 3.11.

3.3.3 Mean Path Length

In a network, the *mean path length* is defined as the average length of the shortest path between two nodes. The mean path length is often called the average path length.

Let D_{ij} be the length of the shortest path between node i and j. In a network, the length of a path is defined using several ways. The shortest path length is often defined as the number of hops between two nodes. The shortest path between two points, called geodesic, is also considered. The definition of the length between two nodes should be symmetrical if the underlying network structure is undirectional. Namely, we impose $D_{ij} = D_{ji}$.

The mean path length is computed as an average of the shortest path length D_{ij} over all the pairs of nodes. For an undirected network of N nodes, the mean path length is defined as

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=j}^{N} D_{ij}, \tag{3.203}$$

or

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=1}^{j-1} D_{ij}, \tag{3.204}$$

where the diagonal components of the shortest path length are zero $D_{ii} = 0$. If two nodes are disconnected, then the path length between them is infinite. As a result, if a network contains disconnected components, then the mean path length l also diverges to infinity. To avoid the divergence of $\langle l \rangle$, we may take the sum over only nodes in the largest connected component. In the case of an undirectional network, we may not find a pair of nodes that has no connections in the largest connected component. However, a directed network often has pairs of nodes that are not reachable. In this case, we can take an average of pairs of nodes in weakly connected components (WCC).

3.3.4 Centrality

3.3.4.1 Degree Centrality

Degree centrality was firstly proposed by Freeman in social network analysis [22]. An underlying idea behind is that nodes which have more links to other nodes may be related to the importance of the nodes. Because they have many links, they may be able to access resources of other nodes. Therefore, we can measure the importance of a node by using its degree.

In an undirected network, nodes differ from one another only in the number of connections which they have. In a directed network, however, it can be important to distinguish centrality based on in-degree with centrality based on out-degree. If a node receives many links (has high in-degree centrality), it is often said to be prominent or to have high prestige. Nodes that have an unusually high out-degree are often said to be influential since they are able to exchange with many others.

Suppose that we have an adjacency matrix A_{ij} for N nodes. In the case of an undirectional network, a degree of the i-th node is calculated using Eq. (3.181). Thus, a degree centrality of the i-th node is defined as

$$C_D(e_i) = \frac{\sum_{j=1}^{N} A_{ij}}{N-1} = \frac{k_i}{N-1}. \tag{3.205}$$

In a directional network, an in-degree of the j-th node is defined as

$$C_D^{(in)}(e_j) = \frac{\sum_{i=1}^{N} A_{ij}}{N-1} = \frac{k_j^{(in)}}{N-1}, \tag{3.206}$$

and the out-degree of the i-th node is defined as

$$C_D^{(out)}(e_i) = \frac{\sum_{j=1}^{N} A_{ij}}{N-1} = \frac{k_i^{(out)}}{N-1}. \tag{3.207}$$

The degree centrality of a point is viewed as important as an index of its communication activity. Obviously, $C_D(e_i)$, $C_D^{(in)}(e_j)$ and $C_D^{(out)}(e_i)$ take a value ranging from 0 to 1.

3.3.4.2 Betweenness Centrality

Betweenness centrality was also formalised by Freeman [21]. It is defined as follows. Suppose that g_{ij} represents the number of geodesic linking e_i and e_j. If we assume that two nodes e_i and e_j are indifferent with respect to which of several alternative geodesic carries their communications, then the probability of using any one of them is

$$\frac{1}{g_{ij}}. \tag{3.208}$$

The potential of point e_k for control of information passing between e_i and e_j may be defined as the probability that e_k falls on a randomly selected geodesic connecting e_i and e_j. By introducing $g_{ij}(e_k)$ to the number of geodesics linking e_i and e_j that contain e_k, we can calculate the probability as

$$b_{ij}(e_k) = \frac{1}{g_{ij}} \times g_{ij}(e_k) = \frac{g_{ij}(e_k)}{g_{ij}}. \tag{3.209}$$

The overall centrality of a node e_k is given by the sum of its partial betweenness values for all unordered pairs of nodes where $i \neq j \neq k$:

$$C_B(e_k) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} b_{ij}(e_k), \tag{3.210}$$

where N is the number of nodes in the network. When e_k falls on the only geodesic connecting a pair of nodes, $C_B(e_k)$ increases by 1. Obviously, $C_B(e_k)$ takes a value ranging from 0 to 1.

3.3.4.3 Eigenvector Centrality

Eigenvector centrality is an index that considers both direct and indirect influences from the other nodes in a network. When thinking about a real network, the importance of a node needs to be quantified by considering the indirect influence to a node that is located more than two links away, the eigenvector centrality is the more practical measure to apply to the real network analysis.

The eigenvector centrality $C_{ev}(e_i)$ of node i in an undirectional network with N nodes can be defined as

$$C_{ev}(e_i) = \frac{1}{\lambda_1(\boldsymbol{A})} \sum_{j=1}^{N} A_{ij} C_{ev}(e_j), \tag{3.211}$$

where A_{ij} represents an element in the i-th row and the j-th column of the adjacency matrix and $\lambda_1(\boldsymbol{A})$ shows the maximum eigenvalue of the adjacency matrix $\boldsymbol{A}$. In the matrix notation, Eq. (3.211) can be expressed as

$$\lambda_1(\boldsymbol{A})\boldsymbol{C}_{ev} = \boldsymbol{A}\boldsymbol{C}_{ev}, \tag{3.212}$$

where $\boldsymbol{C}_{ev}$ is an $N \times 1$ column vector in which each element corresponds to the value of the eigenvalue centrality $C_{ev}(e_i)$.

3.3.4.4 Alpha Centrality

Bonacich [13] proposed an extension of the eigenvector centrality, which is called *alpha centrality*. In the case of an undirectional network, the alpha centrality is defined as

$$\boldsymbol{C}_\alpha = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}\boldsymbol{e}, \tag{3.213}$$

where $\boldsymbol{C}_\alpha$ is an $N \times 1$ column vector in which each element corresponds to the value of the alpha centrality $C_\alpha(e_i)$ of the i-th node. $\boldsymbol{e}$ is an $N \times 1$ column vector in which all elements are 1, and $\boldsymbol{I}$ is a unit matrix. α is an arbitrary parameter that controls the weight of directional or undirectional influence from the other nodes in the network represented by the adjacency matrix $\boldsymbol{A}$. Any symmetrical matrix can be decomposed as follows:

$$\boldsymbol{A}\boldsymbol{X} = \lambda \boldsymbol{X}, \quad \boldsymbol{A} = \boldsymbol{X}\lambda \boldsymbol{X}^{-1} = \boldsymbol{X}\lambda \boldsymbol{X}^{\mathrm{T}}. \tag{3.214}$$

Because the eigenvectors of a symmetrical matrix are orthogonal, the powers of $\boldsymbol{A}$ have similar decompositions.

$$\boldsymbol{A}^k = \sum_{i=1}^{N} \lambda_i^k \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}, \tag{3.215}$$

where $\boldsymbol{v}_i$ is the i-th orthonormal eigenvectors of an $N \times N$ symmetric matrix $\boldsymbol{A}$, which is described as an $N \times 1$ column vector, and λ_i is the associated with the i-th eigenvalues. For $\alpha < 1/\lambda_1$, we have

$$\begin{aligned} \boldsymbol{C}_\alpha &= (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}\boldsymbol{e} = \Big(\sum_{k=0}^{\infty} \alpha^k \sum_{i=1}^{N} \lambda_i^k \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}\Big)\boldsymbol{e} \\ &= \Big(\sum_{i=1}^{N} \big(\sum_{k=0}^{\infty} \alpha^k \lambda_i^k\big) \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}\Big)\boldsymbol{e} = \sum_{i=1}^{N} \frac{\boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}}{1 - \alpha \lambda_i}\boldsymbol{e}. \end{aligned} \tag{3.216}$$

In the case of directional network, the eigenvectors of an asymmetrical adjacency matrix are not orthogonal. Therefore, Eq. (3.213) is a bit different. For an asymmetrical matrix, $\boldsymbol{A}\boldsymbol{X} = \lambda \boldsymbol{X}$, as before, but $\boldsymbol{X}^{\mathrm{T}} \neq \boldsymbol{X}^{-1}$. However, it is true that $\boldsymbol{A} = \boldsymbol{X}\lambda^k \boldsymbol{X}^{-1}$. Letting $\boldsymbol{w}_i$ be the i-th row of $\boldsymbol{X}^{-1}$,

$$\boldsymbol{A} = \sum_{i=1}^{N} \lambda_i \boldsymbol{v}_i \boldsymbol{w}_i, \quad \boldsymbol{A}^k = \sum_{i=1}^{N} \lambda_i^k \boldsymbol{v}_i \boldsymbol{w}_i. \tag{3.217}$$

Therefore,

$$\begin{aligned} \boldsymbol{C}_\alpha &= (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}\boldsymbol{e} = \Big(\sum_{k=0}^{\infty} \alpha^k \sum_{i=1}^{N} \lambda_i^k \boldsymbol{v}_i \boldsymbol{w}_i\Big)\boldsymbol{e} \\ &= \Big(\sum_{i=1}^{N} \big(\sum_{k=0}^{\infty} \alpha^k \lambda_i^k\big) \boldsymbol{v}_i \boldsymbol{w}_i\Big)\boldsymbol{e} = \sum_{i=1}^{N} \frac{\boldsymbol{v}_i \boldsymbol{w}_i}{1 - \alpha \lambda_i}\boldsymbol{e}. \end{aligned} \tag{3.218}$$

If λ_1 is strictly greater than any other eigenvalue, the coefficient for the first term in the final sum in Eq. (3.218) will become more and more dominant as α approaches $1/\lambda_1$.

3.3.5 Network Entropy

The concept of statistical–physical entropy was applied by Bianconi [6, 10] to measure network structure. She considered that the complexity of a network is related to the number of possible configurations of nodes and links under some constraints determined by observations. She calculated the network entropy of an arbitrary undirected network in several cases of constraints. She considered four types of constrains [6, 10]:

1. An ensemble of random networks with a given number of nodes N and links $L = \sum_{i<j} A_{ij}$.
2. An ensemble of networks with given degree sequences $\{k_1, \ldots, k_N\}$ with $k_i = \sum_{j=1} A_{ij}$.
3. An ensemble of networks with given degree sequences $\{k_1, \ldots, k_N\}$ and given average nearest neighbour connectivity $k_{nn}(k) = [\sum_{i,j} \delta(k_i - k) A_{ij} k_j]/(kN_k)$ (with N_k indicating the number of nodes of degree k in the network).
4. An network ensemble with a given degree sequence and a given community structure.

One of the simple cases of network entropy is given by using the number of undirected networks $\mathcal{N}_0 = \frac{N(N-1)}{2} C_L$ with a given number of nodes N and links L as constraints. In this case, the entropy per node can be described as

$$\Sigma_0 = \frac{1}{N} \ln \mathcal{N}_0. \tag{3.219}$$

There is an alternative definition of network entropy to characterise network structure with information-theoretic entropy [16, 30, 36, 42, 45]. Several graph invariants, such as the number of nodes, the node degree sequence and extended degree sequences, have been used in the construction of entropy-based measures [45].

The simplest case of the network entropy is defined by using the node degree. The network entropy is defined as the Shannon entropy for the node degree density. In the case of an undirected network of the number of nodes N with the adjacency matrix A_{ij} $(i, j = 1, \ldots, N)$, the node degree density of the i-th node to the total number of links L is computed as

$$c(e_i) = \frac{\sum_{j=1}^{N} A_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}} = \frac{k_i}{2L}. \tag{3.220}$$

Obviously, $0 \leq c(e_i) \leq 1$ and $\sum_{i=1}^{N} c(e_i) = 1$. Then, the network entropy $H(\boldsymbol{A})$ is defined as

$$H(\boldsymbol{A}) = -\sum_{i=1}^{N} c(e_i) \ln c(e_i). \tag{3.221}$$

The network entropy $H(\boldsymbol{A})$ takes a small value if $c(e_i)$ is heterogeneous and it takes a large value if $c(e_i)$ is homogeneous. The maximum value of $H(\boldsymbol{A})$ is $H_{max} = \ln N$ for a homogeneously connected network with $c(e_i) = 1/N$.

In the case of a directed network, we have two types of network entropies for in-degree and out-degree. They are defined as

$$H^{(in)}(\boldsymbol{A}) = -\sum_{j=1}^{N} c^{(in)}(e_j) \ln c^{(in)}(e_j), \tag{3.222}$$

$$H^{(out)}(\boldsymbol{A}) = -\sum_{i=1}^{N} c^{(out)}(e_i) \ln c^{(out)}(e_i), \tag{3.223}$$

where

$$c^{(in)}(e_j) = \frac{\sum_{i=1}^{N} A_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}} = \frac{k_j^{(in)}}{L}, \quad c^{(out)}(e_i) = \frac{\sum_{j=1}^{N} A_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}} = \frac{k_i^{(out)}}{L}. \tag{3.224}$$

Note that $0 \leq c^{(in)}(e_j) \leq 1$, $\sum_{j=1}^{N} c^{(in)}(e_j) = 1$, $0 \leq c^{(out)}(e_i) \leq 1$ and $\sum_{i=1}^{N} c^{(out)}(e_i) = 1$.

3.3.6 Assortativity Coefficient

The assortative coefficient is used to measure the level of homophily of the graph. Suppose that there are K types of nodes in a network. If the same type of nodes tends to be connected to each other, then this is defined as an *assortative* network. If the different types of nodes to be connected to each other, then this is defined as a *disassortative* network. Newman defines two kinds of assortativity coefficients [31, 32].

Let q_{ij} be the fraction of links in a network that connect a node of type i to one of type j. For an undirectional network, obviously we have $q_{ij} = q_{ji}$. It satisfies the sum rules

$$\sum_{i=1}^{K}\sum_{j=1}^{K} q_{ij} = 1, \quad \sum_{j=1}^{K} q_{ij} = a_i, \quad \sum_{i=1}^{K} q_{ij} = b_j, \tag{3.225}$$

where a_i and b_j are the fraction of each type of end of a link that is connected to nodes of type i and j, respectively. On an undirectional network, $a_i = b_i$ holds.

The first assortative coefficient [31] is defined as

$$r = \frac{\sum_{i=1}^{K} q_{ii} - \sum_{i=1}^{K} a_i b_i}{1 - \sum_{i=1}^{K} a_i b_i}. \tag{3.226}$$

If Eq. (3.226) gives $r = 0$, then there is no assortative mixing in the network. $r = 1$ implies that there is perfect assortative mixing and $\sum_{i=1}^{K} q_{ii} = 1$. If the network is perfectly, disassortative so that every link connects two nodes of different types, then r is negative and takes the minimum value $-(\sum_{i=1}^{K} a_i b_i)/(1 - \sum_{i=1}^{K} a_i b_i)$.

The second assortativity variant is based on values assigned to the nodes [32]. This is defined as

$$r = \frac{\sum_{j=1}^{K}\sum_{k=1}^{K} jk(q_{jk} - b_j b_k)}{\sum_{k=1}^{K} k^2 b_k - (\sum_{k=1}^{K} k b_k)^2}, \tag{3.227}$$

which is equivalent to the Pearson correlation coefficient of the degrees at either ends of a link and lies in the range $-1 \leq r \leq 1$. In the case of directed network, the assortativity takes different values for in-degree and out-degree.

3.3.7 Community Detection

Community detection in networks is an important issue in complex network studies, and many different methods are proposed [20]. The simple method to detect communities in a network is based on the maximisation of modularity measure [33], defined as

$$Q = \sum_{s=1}^{n_M} \left\{ \frac{L_s}{L} - \left(\frac{N_s}{2L} \right)^2 \right\}, \tag{3.228}$$

where n_M denote the number of communities (modules), N_s and L_s represent the number of nodes in community s and the number of links inside community s. The first term of the summand in Eq. (3.228) is the fraction of links inside community s and the second term represents the expected fraction of links in that community if links were located at random in the network. The number of communities n_M is also a variable of which value is obtained by the maximisation.

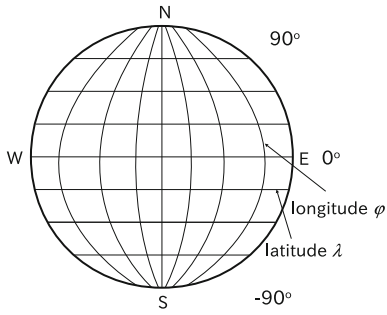


Fig. 3.13 A conceptual illustration of geographic coordinate system

3.4 Spatial Analysis

Spatial analysis and geographic information system (GIS) provide us with the ability to analyse data with geographical information. In spatial analysis, we treat data of physical quantities with geographical positions. Visualisation, density estimation and computation of spatial statistics are often used. In this section, we will address the fundamental methods for spatial descriptions and spatial autocorrelations. Readers may find useful information in a number of books on spatial analysis [11, 19].

3.4.1 Geographic Coordinate System

A geographic coordinate system enables every location on Earth to be specified by the latitude, longitude and elevation. Figure 3.13 shows a conceptual illustration of geographic coordinate system. The latitude is defined from north to south. The longitude is defined from east to west.

The latitude of a point on Earth's surface is the angle between the equatorial plane and a curve that passes through that point and the centre of Earth to the surface of a reference ellipsoid that approximates the shape of Earth. This line passes from the centre of Earth except at the poles and the equator where it passes through Earth's centre. Curves joining points of the same latitude trace circles on the surface of Earth, called parallels, as they are parallel to the equator, and to each other.

The longitude of a point on Earth's surface is the angle east or west from a reference meridian to another meridian that passes through that point. A curve passing near the Royal Observatory in Greenwich (UK) has been chosen as the international zero-longitude reference curve, which is called the Prime Meridian. Places to the east are in the eastern hemisphere, and places to the west are in the western hemisphere.

Let ϕ° and λ° be the latitude and longitude at position s, respectively. We assume that they are measured in degrees. The latitude ϕ° takes a value ranging from -90° to 90°. The north pole is 90° and the south pole is -90°. The 0° parallel of latitude corresponds to the equator. The longitude λ° takes a value from -180° to 180°. Thus,

the western hemisphere ranges from $-180°$ to $0°$, and the eastern hemisphere from $0°$ to $180°$. The antipodal meridian of Greenwich is both $-180°$ and $180°$.

If we measure the latitude and longitude in radians, the longitude λ takes a value ranging from $-\pi$ to π and the latitude ϕ from $-\pi/2$ to $\pi/2$. The equality between radians and degrees can be used in the interpretation.

$$\lambda = \pi\lambda^\circ/180, \quad \phi = \pi\phi^\circ/180. \tag{3.229}$$

3.4.2 Data on Geography

Grid data of Earth's surface can be downloaded from NOAA Data Centre.[1] ETOPO1 [3] is a 1 arc-minute global relief model of Earth's surface that integrates land topography and ocean bathymetry. It is built from numerous global and regional data sets. There are two versions of the Grid in ETOPO1: Ice Surface and Bedrock.

Shape files are the most popular data used in geostatistics. GADM is a spatial database of the location of the world's administrative area for use in GIS.[2] This contains various attributes ('*spatial features*'), such as the name and variant names for each Global Administrative Area.

3.4.3 Map Projections

In this section, we address how to transform the latitude and the longitude into horizontal and vertical positions. This is called a *map projection*, which provides a systematic method to transform the latitudes and longitudes of locations on the surface of a sphere or an ellipsoid into locations on a plane. There are several map projections classified into:

- Cylindrical type
- Pseudocylindrical type
- Azimuthal type
- Pseudoazimuthal type
- Conic type
- Pseudoconical type
- Polyhedral type
- Retroazimuthal type

We measure both the latitude ϕ ($-\pi/2 \leq \phi \leq \pi/2$) and longitude λ ($-\pi \leq \lambda \leq \pi$) in radians. Let x and y be the horizontal and vertical positions. We see, for example, four typical map projections in cylindrical and pseudcylindrical types below and

[1] http://www.ngdc.noaa.gov/mgg/global.

[2] GADM database: www.gadm.org.

draw maps with actual data. We address four types of typical projections in the list above.

3.4.3.1 Equirectangular Projection

Equirectangular projection is a simple map projection, which is classified into cylindrical projections. This is defined as:

$$\begin{cases} x = \left\{(\lambda - \lambda_0 + \pi) \bmod 2\pi\right\} - \pi \\ y = \phi \end{cases}, \tag{3.230}$$

where the point $(0, 0)$ is at the centre of the resulting projection. This position is mapped into $\lambda = \lambda_0$ (λ_0 is the central meridian), and $\phi = 0$. Figure 3.14a shows the world map drawn using equirectangular projection for $\lambda_0 = \frac{135}{180}\pi$. The range of x is from $-\pi$ to π, and of y from $-\pi/2$ to $\pi/2$.

3.4.3.2 Lambert Cylindrical Equal-Area Projection

Lambert cylindrical equal-area projection is a cylindrical, equal-area map projection. It is defined as

$$\begin{cases} x = \left\{(\lambda - \lambda_0 + \pi) \bmod 2\pi\right\} - \pi \\ y = \sin\phi \end{cases}, \tag{3.231}$$

where λ_0 is the central meridian. Figure 3.14b shows the world map drawn using the Lambert cylindrical equal-area projection for $\lambda_0 = \frac{135}{180}\pi$. The range of x is from $-\pi$ to π and of y from -1 to 1.

3.4.3.3 Sinusoidal Projection

The sinusoidal projection is classified into a pseudocylindrical equal-area map projection. This is also called Sanson projection. This projection is described as

$$\begin{cases} x = \left[\left\{(\lambda - \lambda_0 + \pi) \bmod 2\pi\right\} - \pi\right] \times \cos\phi \\ y = \phi \end{cases}, \tag{3.232}$$

where λ_0 is the central meridian. Figure 3.14c shows the world map drawn using the sinusoidal projection for $\lambda_0 = \frac{135}{180}\pi$. The range of x is from $-\pi$ to π and of y from $-\pi/2$ to $\pi/2$.

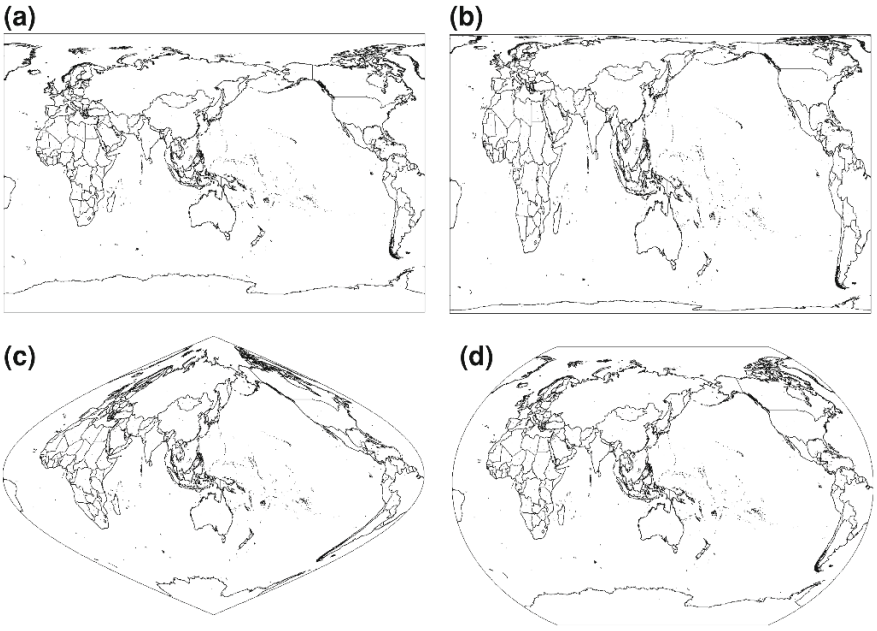


Fig. 3.14 **a** The world map drawn with the Equirectangular projection. **b** The world map drawn with the Lambert cylindrical equal-area projection. **c** The world map drawn with the sinusoidal projection. **d** The world map drawn with the WagnerVI projection. These pictures are drawn with Grid data of Earth's surface (ETOPO1 Ice Surface Global Relief Model) downloaded from NOAA Data Centre

3.4.3.4 Wagner VI Projection

Wagner VI projection is classified into a pseudocylindrical and compromise map projection, which is defined as:

$$\begin{cases} x = \left[\{(\lambda - \lambda_0 + \pi) \bmod 2\pi\} - \pi \right] \times \sqrt{1 - 3\left(\frac{\phi}{\pi}\right)^2} \\ y = \phi \end{cases}, \tag{3.233}$$

Figure 3.14d shows the world map drawn using Wagner VI projection for $\lambda_0 = \frac{135}{180}\pi$.

3.4.4 Geodesic Distance

There are several types of approximation to calculate geodesic distance from two points of latitude and longitude. The simplest method is an approximation under the assumption that Earth is a sphere. This approximation is referred to as Vincenty's formulae. Let ϕ_s, λ_s, ϕ_f and λ_f be the geographical latitude and longitude of two

points s and f. The distance D between s and f is given by

$$D = r \tan^{-1} \left(\frac{\sqrt{(\cos\phi_s \sin\Delta\lambda)^2 + (\cos\phi_s \sin\phi_f - \sin\phi_s \cos\phi_f \cos\Delta\lambda)^2}}{\sin\phi_s \sin\phi_f + \cos\phi_s \cos\phi_f \cos\Delta\lambda} \right), \tag{3.234}$$

where r represents the earth's radius ($r = 6,371.2$ km) and $\Delta\lambda = \lambda_s - \lambda_f$.

More accurately, the shape of Earth closely resembles a flattened sphere (ellipsoid) with an equatorial radius a of 6,378.137 km, and the polar semi-minor axis b equals 6,356.7523142 km. Therefore, the flattening of Earth, which is defined as

$$\text{flattening} = f = \frac{a - b}{a}, \tag{3.235}$$

is calculated as $f = 1/298.257223563$. This model of Earth is employed in the World Geodetic System 1984 (WGS 84), which is meant to be Earth's centre of mass. The error of distance measured by the WGS 84 is believed to be less than 2 cm.

3.4.5 Spatial Autocorrelation

To estimate spatial autocorrelation statistics, classically, there are two types of spatial autocorrelation statistics, called Moran's [29] I and Geary's C [24]. The local spatial autocorrelation called Getis-Ord's G [34] is also considered. The spatial data is described as a value, which is called "*feature*", with information on its location.

To compute spatial autocorrelation, we first need to define how to measure the proximity between two observations. Namely, the distance measure must be introduced. These distances are expressed as a matrix called the weight matrix, which defines the relationships between locations where measurements were made. If we have T observations collected at each location, then the weight matrix will be a $T \times T$ matrix with zeros on the diagonal. Examples of the weight matrix are as follows:

- The weight for any two different locations is a constant.
- All observations within a specified distance have a fixed weight.
- The nearest neighbours up to K-distance have a fixed weight, and all others are zero.
- The weight is proportional to the inverse distance, the inverse distance squared or the inverse distance up to a specified distance.

3.4.5.1 Moran's *I*

Moran's I assumes that the null hypothesis states that there is no spatial clustering of the values associated with the geographic features in the study area [29]. When the p-value is small, the null hypothesis can be rejected. Namely, it indicates that there is a spatial clustering in the features. If the index value I is greater than 0, the set of features exhibits a clustered pattern. Otherwise, the set of features shows a dispersed pattern.

Suppose that we have T features x_i with the geographical position. The Moran's I statistic for spatial autocorrelation is defined as

$$I = \frac{T}{S_0} \frac{\sum_{i=1}^{T} \sum_{j=1}^{T} w_{ij}(x_i - \bar{X})(x_j - \bar{X})}{\sum_{j=1}^{T}(x_j - \bar{X})^2}, \tag{3.236}$$

where $\bar{X}$ represents the empirical mean defined as $\bar{X} = \sum_{i=1}^{T} x_i / T$, w_{ij} is the spatial weight between feature i and j and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^{T} \sum_{j=1}^{T} w_{ij}. \tag{3.237}$$

Negative (positive) values of Moran's I indicate negative (positive) spatial correlation. Since the mean and the variance of Moran's I for the null hypothesis are given as

$$\mathrm{E}[I] = \frac{-1}{T-1}, \tag{3.238}$$

$$\begin{aligned} \mathrm{Var}[I] &= \mathrm{E}[I^2] - \mathrm{E}[I]^2 \\ &= \frac{T S_4 - S_3 S_5}{(T-1)(T-2)(T-3) S_0^2} - \frac{1}{(T-1)^2}, \end{aligned} \tag{3.239}$$

where

$$S_1 = \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{T} (w_{ij} + w_{ji})^2, \tag{3.240}$$

$$S_2 = \sum_{i=1}^{T} \left(\sum_{j=1}^{T} w_{ij} + \sum_{j=1}^{T} w_{ji} \right)^2, \tag{3.241}$$

$$S_3 = \frac{\left(\sum_{i=1}^{T} (x_i - \bar{X}) \right)^4}{\left(\sum_{i=1}^{T} (x_i - \bar{X})^2 \right)^2}, \tag{3.242}$$

$$S_4 = (T^2 - 3T + 3)S_1 - TS_2 + 3S_0^2, \tag{3.243}$$
$$S_5 = T(T-1)S_1 - 2TS_2 + 6S_0^2, \tag{3.244}$$

the p-value for the null hypothesis is written as

$$p = \mathrm{erfc}\left(\frac{|I - \mathrm{E}[I]|}{\sqrt{2\mathrm{Var}[I]}}\right). \tag{3.245}$$

$z = \frac{I - \mathrm{E}[I]}{\sqrt{\mathrm{Var}[I]}}$ is called z-score, which is used to confirm the significance level of the spatial autocorrelation.

3.4.5.2 Geary's C

While Moran's I measures global spatial autocorrelation, *Geary's C* is more sensitive to local spatial autocorrelation [24]. Geary's C is related to Moran's I, but it is not identical.

Suppose that we have T features x_i with the geographical position. Geary's C is defined as

$$C = \frac{(T-1)\sum_{i=1}^{T}\sum_{j=1}^{T} w_{ij}(x_i - x_j)^2}{2S_0 \sum_{i=1}^{T}(x_i - \bar{X})^2}, \tag{3.246}$$

where $\bar{X}$ represents the empirical mean of x_i, w_{ij} a weight matrix, and S_0 the sum of all w_{ij} defined in Eq. (3.237). Geary's C takes a non-negative value. Since the expectation value of the null hypothesis (the absence of spatial autocorrelation) independently of the specific weight matrix, the value of C less than 1 implies positive spatial autocorrelation. The negative spatial autocorrelation corresponds to the value of C more than 1. The variance of Geary's C is given as

$$\mathrm{E}[C] = 1, \tag{3.247}$$
$$\begin{aligned}\mathrm{Var}[C] = {} & \frac{1}{T(T-2)(T-3)S_0^2} \\ & \times \Big\{ S_0^2[T^2 - 3 - (T-1)^2 S_3] \\ & + S_1(T-1)[T^2 - 3T + 3 - (T-1)S_3] \\ & + \frac{1}{4}S_2(T-1)[S_3(T^2 - T + 2) - (T^2 + 3T - 6)] \Big\},\end{aligned} \tag{3.248}$$

where S_0, S_1, S_2 and S_3 are computed from Eqs. (3.237) to (3.242). The p-value is computed from Eq. (3.245) as well as Moran's I. Consequently, the p-value for the null hypothesis is given as

$$p = \mathrm{erfc}\left(\frac{|C - \mathrm{E}[I]|}{\sqrt{2\mathrm{Var}[C]}}\right). \tag{3.249}$$

$z = \frac{C-\mathrm{E}[C]}{\sqrt{\mathrm{Var}[C]}}$ is called z-score, which is used to confirm the significance level of the spatial autocorrelation.

3.4.5.3 Local Moran's I Statistic

Equation (3.236) estimates a global tendency of clustering but does not measure a local spatial autocorrelation. To measure a local spatial autocorrelation, the local Moran's I statistic is defined as

$$I_i = \frac{T}{S_0}\frac{(x_i - \bar{X})\sum_{j=1}^{T} w_{ij}(x_j - \bar{X})}{\sum_{j=1}^{T}(x_j - \bar{X})^2}, \tag{3.250}$$

where $I = \sum_{i=1}^{T} I_i$ holds. I_i can be used to measure the local tendency of clustering. According to Anselin [7], Moran's I is interpreted as a regression coefficient of the linear association between $z_i = x_i - \bar{X}$ and $\sum_{j=1}^{T} w_{ij} z_j$. From Eq. (3.57), for OLS regression of $\sum_{j=1}^{T} w_{ij} z_j = a z_i + b$, we can derive

$$\begin{aligned}
\hat{a} &= \frac{\sum_{j=1}^{T} z_j \sum_{k=1}^{T} w_{jk} z_k - (\sum_{i=1}^{T} z_i)(\sum_{j=1}^{T}\sum_{k=1}^{T} w_{jk} z_k)}{\sum_{j=1}^{T} z_j^2} \\
&= \frac{\sum_{j=1}^{T}\sum_{k=1}^{T} z_j w_{jk} z_k}{\sum_{j=1}^{T} z_j^2}, \\
&= \frac{\sum_{j=1}^{T}\sum_{k=1}^{T}(x_j - \bar{X}) w_{jk} (x_k - \bar{X})}{\sum_{j=1}^{T}(x_j - \bar{X})^2} = \frac{S_0}{T} I,
\end{aligned} \tag{3.251}$$

where $\sum_{i=1}^{T} z_i = 0$ is used. This also forms a method to visualise a tendency of local spatial autocorrelations from a bivariate scatter plot of $\sum_{j=1}^{T} w_{ij} x_j$ against x_i. This is called *univariate Moran scatter plot*. If we consider the Moran scatter plots for different features x_i and y_i and make scatter plots of $\sum_{j=1}^{T} w_{ij} y_j$ in terms of x_i, then this is called *bivariate Moran scatter plot*.

3.4.5.4 Getis-Ord's Local G

Getis and Ord introduced a family of statistics, G, that can be used as measures of spatial association in a number of circumstances [34].

Suppose that we have T features x_i with the geographical position. Getis-Ord's Local G is defined as

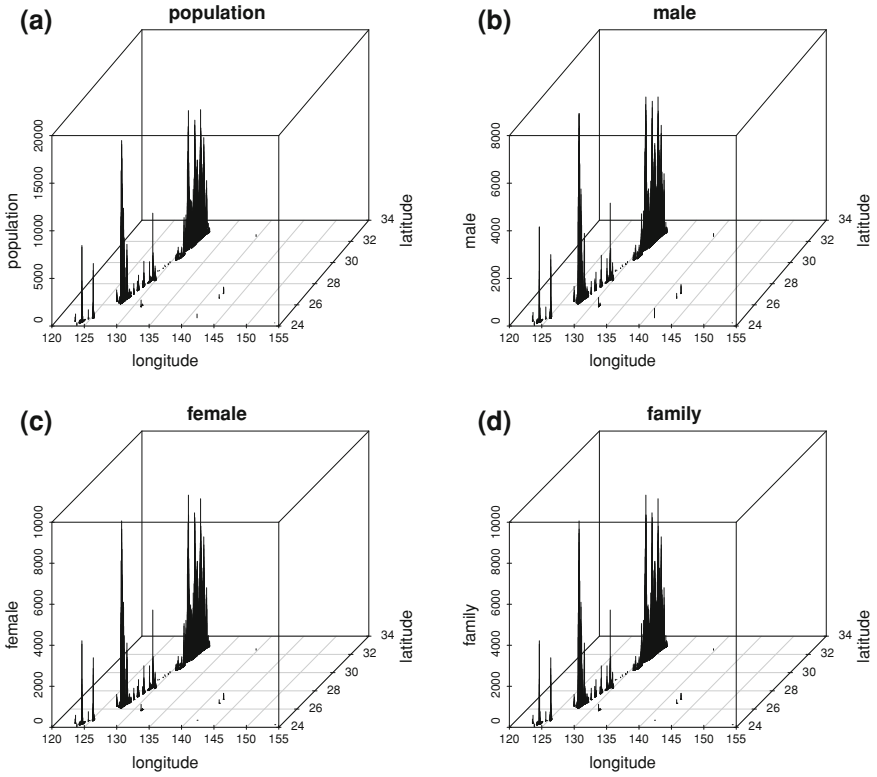


Fig. 3.15 **a** 1 km square statistics of the Japanese census population in 2010, **b** the number of males, **c** the number of females and **d** the number of families

$$G_i(d) = \frac{\sum_{j=1}^{T} w_{ij}(d)(x_j - \bar{X})}{\sum_{j=1}^{T}(x_j - \bar{X})}, \tag{3.252}$$

where $\bar{X}$ represents the empirical mean of x_i and $w_{ij}(d)$ is a binary, symmetrical weight matrix with 1 for all points j within distance d of point i and 0 otherwise. T is equal to the total number of features. Essentially, the G_i statistic represents a ratio of the values within distance d of point i to the sum of all values minus the value at point i. G_i^* is computed from the randomised data using the bootstrap method or the jackknife method to evaluate the significance level of the spacial autocorrelation. G_i and G_i^* enable us to detect pockets of spatial association that may not be evident when using Moran's I and Geary's C.

Table 3.3 The values of spatial autocorrelation for Moran's I and Geary's C; data for 1 km square statistics for the Japanese population

Type	Moran's I	E[I]	$\sqrt{\text{Var}[I]}$	p-value
Population	0.729504	−0.000050	0.000017	0.000000
Male	0.730892	−0.000050	0.000017	0.000000
Female	0.726565	−0.000050	0.000017	0.000000
Family	0.732767	−0.000050	0.000017	0.000000
Type	Geary's C	E[C]	$\sqrt{\text{Var}[C]}$	p-value
Population	0.272013	1.000000	0.000044	0.000000
Male	0.270554	1.000000	0.000043	0.000000
Female	0.275036	1.000000	0.000044	0.000000
Family	0.269214	1.000000	0.000054	0.000000

3.4.5.5 Empirical Analysis

Using 1 km square statistics data for the Japanese population as spatial data, let us consider an example of empirical analysis for spatial autocorrelation. The data were generated based on Japanese Census, by the Statistics Bureau of the Ministry of Internal Affairs and Communications.[3] In the dataset, four types of 1 km square statistics are included: population, male population, female population, and the number of families. Figure 3.15 shows the spatial distribution of the Japanese population of the southern part of Japan.

Moran's I and Geary's C are computed by using R for statistical computing. The commands **moran.test**() and **geary.test**() in the library '**spdep**' are available to calculate Moran's I and Geary's C. The commands **moran.test**() and **geary.test**() also need a list of neighbours, which is constructed by the command **tri2nb**(). Table 3.3 shows the results of spatial autocorrelations for both cases. In the four cases, the values of the Moran's I are positive with statistical significance. The values of the Geary's C are also less than 1. Since, for all cases, I and C show positive spatial autocorrelation and the p-value is zero, the null hypothesis that the population is randomly distributed is rejected with statistical significance. Furthermore, values of the local Moran's I are computed by using **localmoran**() and the univariate Moran scatter plot is drawn by using **moran.plot**(). Figure 3.16 shows the univariate Moran scatter plots for population. We can see some outliers in the local Moran autocorrelation. For example, the grid #581 corresponds to the place (latitude, longitude) = (26.2, 127.6875). This is the centre place of Naha city.

[3] http://www.stat.go.jp/english/index.htm.

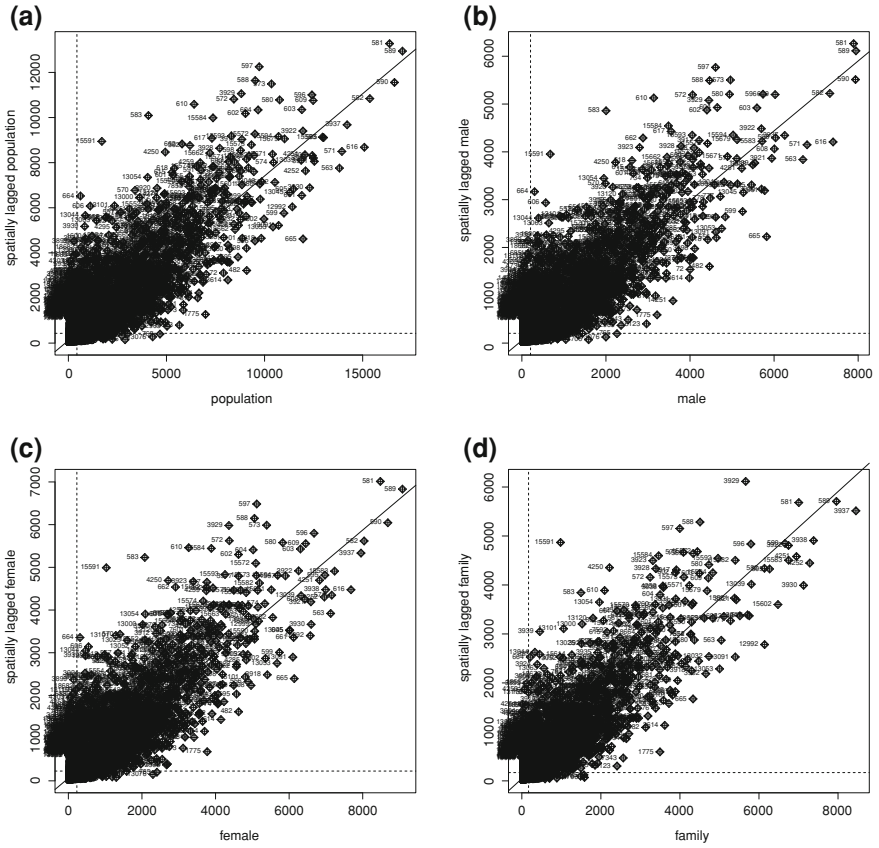


Fig. 3.16 The univariate Moran scatter plots for **a** population, **b** male, **c** female and **d** family in 20,000 1-km square grids of a southern part of Japan

3.5 Combinations of Methods

We may consider combination of methods mentioned below for data analysis. We can use methods of time series analysis and spatial analysis at the same time, which is mentioned as spatial temporal analysis [4, 17]. Furthermore, we can consider network analysis with geographical information, which is called spatial network analysis [9].

For example, let us consider a Japanese domestic air transportation network which were shown in Fig. 3.7. In fact, the network structure is embedded in space. Figure 3.17 shows a projection of the Japanese domestic air traffic network on the map. This can be drawn as a spatially embedded network. We can see that there are many flights between large airports (such as Haneda Airport, Osaka Airport, Fukuoka Airport, New Chitose Airport and so on). The location of large airports seems to have a correlation with population density. Figure 3.17 shows 1 km square statistics of the

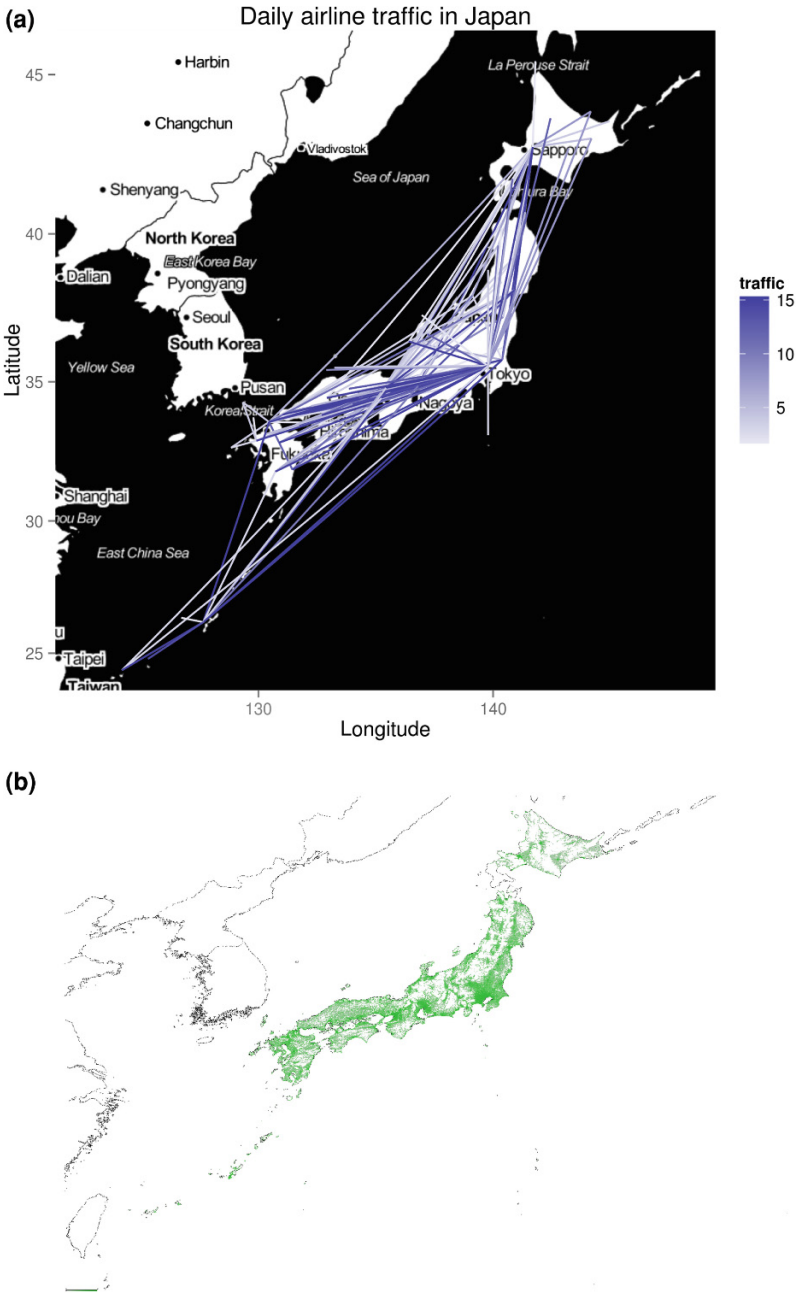


Fig. 3.17 **a** The Japanese domestic air traffic network on a map. The colour represents the daily number of connections between two airports. *Blue colour* corresponds to the number of flights. **b** The population density of Japan in 2010. *Green colour* corresponds to the number of census population

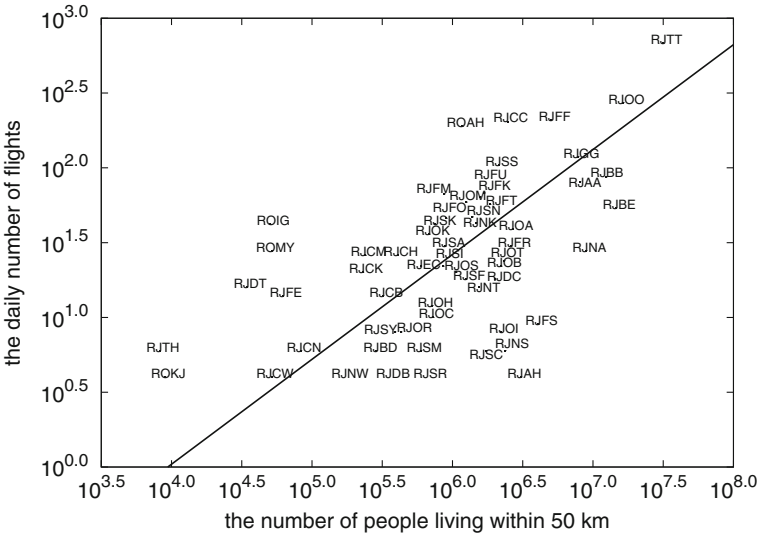


Fig. 3.18 The relationship between the daily number of flights both departing from and arriving at the airports on 15 October 2013 and the number of people who live within a radius of 50 km from the airport in 2010. The *solid line* represents a power-law relationship: (the daily number of flights) $=$ C(the number of people living within 50 km)^a, where a is estimated as 0.701124 (the standard error 0.086132) and $\log_{10} C = -2.785608$ (the standard error 0.516015)

Japanese census population in 2010. It is found that the large airports are located at neighbour of large cities. In order to confirm this intuitive understanding, we regard the relationship between the daily number of flights both departing from and arriving at the airports (this is a degree of network analysis $k_i^{(in)} + k_i^{(out)}$) and population (the number of people who live within a radius of 50 km from the airport). Figure 3.18 shows the relationship in double logarithmic plots. We can see an positive correlation between the population and the number of flights. To evaluate the relationship, we assume the power-law relationship:

$$\text{(the daily number of flights)} = C \times \text{(the number of people living within 50 km)}^a, \tag{3.253}$$

where a represents a power-law exponent and C is a positive constant. By using the RMA regression for the common logarithm for relationships, we obtained parameter estimates as $a = 0.701124(0.086132)$ and $\log_{10} C = -2.785608(0.516015)$. This means that the number of flights landing from and taking off at the airport increases $10^{0.701124} \approx 5$ times if the number of people who live within a radius of 50 km from the airport increases 10 times.

Appendix A: Proof of 0 ln 0

Let us consider

$$h = \lim_{x \to +0} x \ln x. \tag{3.254}$$

Putting $x = e^{-z}$ one has

$$h = -\lim_{x \to +0} z e^{-z}. \tag{3.255}$$

By using the Taylor expansion of $e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k$, one obtains

$$\begin{aligned} h &= -\lim_{z \to \infty} z e^{-z} \\ &= -\lim_{z \to \infty} z / e^z \\ &= -\lim_{z \to \infty} \frac{z}{\sum_{k=0} \frac{1}{k!} z^k} \\ &= -\lim_{z \to \infty} \frac{1}{\sum_{k=0} \frac{1}{k!} z^{k+1}} \\ &= 0. \end{aligned} \tag{3.256}$$

Therefore, we gets

$$h = \lim_{x \to 0} x \ln x = 0 \ln 0 = 0. \tag{3.257}$$

Appendix B: Derivation of the Mean Square Error of RMA Regression

The mean square error MSE of the RMA regression is defined as

$$MSE = \frac{1}{T-2} \sum_{i=1}^{T} (y_i - \hat{a} x_i - \hat{b})^2. \tag{3.258}$$

Inserting Eq. (3.70) into Eq. (3.258), we get

$$
\begin{aligned}
MSE &= \frac{1}{T-2}\sum_{i=1}^{T}(y_i - \hat{a}x_i - \hat{b})^2 \\
&= \frac{1}{T-2}\sum_{i=1}^{T}\left\{y_i - \hat{a}x_i - \left(\frac{\sum_{i=1}^{T} y_i}{T} - \hat{a}\frac{\sum_{i=1}^{T} x_i}{T}\right)\right\}^2 \\
&= \frac{1}{T-2}\sum_{i=1}^{T}\left\{\left(y_i - \frac{\sum_{i=1}^{T} y_i}{T}\right) - \hat{a}\left(x_i - \frac{\sum_{i=1}^{T} x_i}{T}\right)\right\}^2 \\
&= \frac{1}{T-2}\sum_{i=1}^{T}\left\{\left(y_i - \frac{\sum_{i=1}^{T} y_i}{T}\right)^2 + \hat{a}^2\left(x_i - \frac{\sum_{i=1}^{T} x_i}{T}\right)^2\right. \\
&\quad \left. - 2\hat{a}\left(x_i - \frac{\sum_{i=1}^{T} x_i}{T}\right)\left(y_i - \frac{\sum_{i=1}^{T} y_i}{T}\right)\right\}. \qquad (3.259)
\end{aligned}
$$

This is also written as

$$
MSE = \frac{T}{T-2}\Big\{\mathrm{Var}[Y] + \hat{a}^2\mathrm{Var}[X] - 2\hat{a}\mathrm{Cov}[X, Y]\Big\}. \qquad (3.260)
$$

Inserting Eq. (3.74) into Eq. (3.260), consequently we obtain

$$
MSE = \Big(\mathrm{Var}[Y] - \hat{a}\mathrm{Cov}[X, Y]\Big)\frac{2T}{T-2}. \qquad (3.261)
$$

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (eds.) Proceeding of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
3. Amante, C., Eakins, B.W.: ETOPO1 1 Arc-Minute Global Relief Model: procedures, data sources and analysis. NOAA Technical Memorandum NESDIS NGDC-24, 19 Mar 2009
4. Andrienko, G., Andrienko, N.: Exploratory Analysis of Spatial and Temporal Data- A Systematic Approach. Springer, Berlin (2006)
5. Anderson, T.W., Darling, D.A.: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Stat. **23**, 193–212 (1952)
6. Anand, K., Bianconi, G.: Entropy measures for networks- Toward an information theory of complex topologies. Phys. Rev. E **80**, 045102 (2009)
7. Anselin, L.: The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer, M., Scholten, H., Unwin, D. (eds.) Spatial Analytical Perspectives on GIS, pp. 111–125. Taylor and Francis, London (1996)
8. Barrat, A., Barthlemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge (2008)
9. Barthélemy, M.: Spatial networks. Phys. Rep. **499**(1–3), 1–101 (2011)
10. Bianconi, G.: Entropy of network ensembles. Phys. Rev. E **79**, 036114 (2009)

11. Bivand, R.S., Gómez-Rubio, V., Pebesma, E.: Applied Spatial Data Analysis with R, 2nd edn. Springer, New York (2013)
12. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks—structure and dynamics. Phys. Rep. **424**(4–5), 175–308 (2005). http://dx.doi.org/10.1016/j.physrep.2005.10.009
13. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. **23**, 191–201 (2001)
14. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT'2010, pp. 177–186. Springer, Berlin (2010)
15. Box, G.E.P., Jenkins, G.M.: Time Series Analysis, Forcasting and Control. Holden-Day, San Francisco (1976)
16. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Inf. Sci. **181**, 57–78 (2011)
17. Eshel, G.: Spatiotemporal Data Analysis. Princeton University Press, Princeton (2012)
18. Efron, B.: Bootstrap methods—another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)
19. Fischer, M.M., Getis, A. (eds.): Handbook of Applied Spatial Analysis Software Tools, Methods and Applications. Springer, Hidelberg (2011)
20. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
21. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
22. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**, 215–239 (1978/1979)
23. Friedman, J., Bohonak, A.J., Levine, R.A.: When are two pieces better than one—fitting and testing OLS and RMA regressions. Environmetrics **24**, 306–316 (2013)
24. Geary, R.C.: The contiguity ratio and statistical mapping. Inc. Stat. **5**(3), 115–127+129–146 (1954)
25. Huang, W., Zhang, Y.: Estimating structural change in linear simultaneous equations. In: Econometric Society 2004 Australasian Meetings 110, Econometric Society (2004)
26. Kariya, T., Kurata, H.: Generalized Least Squares. Wiley, Chichester (2004)
27. Knight, F.H.: Risk, Uncertainty and Profit. Houghton Mifflin, New York (1921)
28. Lin, J.: Divergence measures base on the Shannon entropy. IEEE Trans. Info. Theor. **37**, 145–151 (1991)
29. Moran, P.A.P: Notes on continuous stochastic phenomena. Biometrika **37**, 17–23 (1950)
30. Mowshowitz, A.: Entropy and the complexity of graphs- I. An index of the relative complexity of a graph. Bull. Math. Biophys. **30**, 175–204 (1968)
31. Newman, M.E.J.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003)
32. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
33. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
34. Ord, J.K., Getis, A.: Local spatial autocorrelation statistics—distributional issues and an application. Geogr. Anal. **27**(4), 286–306 (1995)
35. Quenouille, M.H.: Approximate tests of correlation in time-series. J. R. Statist. Soc. B **11**, 68–84 (1949)
36. Rashevsky, N.: Life, information theory, and topology. Bull. Math. Biophys. **17**, 229–235 (1955)
37. Ricker, W.E.: Linear regressions in fishery research. J. Fish. Res. Board Can. **30**, 409–434 (1973)
38. Shao, J., Tu, D.: The Jackknife and Bootstrap. Springer, New York (1995)
39. Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. **19**, 279–281 (1948)
40. Smith, R.J.: Use and misuse of the reduced major axis for line-fitting. Am. J. Phys. Anthropol. **140**(3), 476–486 (2009)
41. Sokal, R.R., Rohlf, F.J.: Biometry—the principles and practice of statistics in biological research. W. H. Freemann & Company, New York (1995)

42. Trucco, E.: A note on the information content of graphs. Bull. Math. Biophys. **18**, 129–135 (1956)
43. Watts, D.J.: Small Worlds—The Dynamics of Networks Between Order and Randomness. Princeton University Press, Princeton (1999)
44. Weisberg, S.: Applied Linear Regression. Wiley, Hoboken (2005)
45. Wilhelm, T., Hollunder, J.: Information theoretic description of networks. Physica A **385**, 385–396 (2007)
46. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**(1), 60–62 (1938)

Chapter 4
Data in Computers

Abstract The applied data-centric social sciences are cyber-enabled and require the use of inductive strategies to define problems and challenges. Thus, we require the use of computers to process a large number of data points. In this chapter, we will see how computers can be used to acquire, handle and analyse data.

4.1 Computers and Data

Figure 4.1 conceptually illustrates the pipeline from data acquisition to data analysis. Data is acquired from data providers, and stored as some files of certain types on computers. In order to easily access the data, a database server is normally used. Analysis software gets data from the database server with some filters and computes statistics or visualises the data.

The implementation of program codes to conduct data analysis and processing by combining several computer languages enables us to conduct statistical analysis with high reliability for a short development period. Computer software has evolved in a very specific way to cohere with data management requirements. Three types of software are used in order to collect, process and analyse data: databases (MySQL, PostgreSQL, Oracle, MangoDB, Shunsaku, TX1, Hadoop, and so on), script languages (Perl, Ruby, PHP, Python and so on), and analytics languages (R, S-plus, Matlab, Octave, and so on).

The script languages are used to control other computer programs and process data. The Structured Query Language (SQL) of databases is used to insert, extract and sort data under certain constraints. The analytics languages are used to compute several statistics and parameter estimates of a model and conduct regression analysis.

Since the data-centric science is cyber-enabled and requires the use of inductive methodologies, certain functional elements related to both the hardware and the software of computers are mandatory. In this chapter, we will explore how a research environment conductive to the applied data-centric social sciences can be set up.

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_4, © Springer Japan 2014

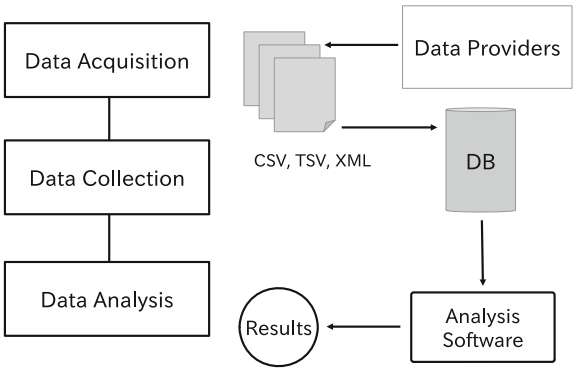


Fig. 4.1 A schematic illustration of a pipeline from data acquisition to data analysis

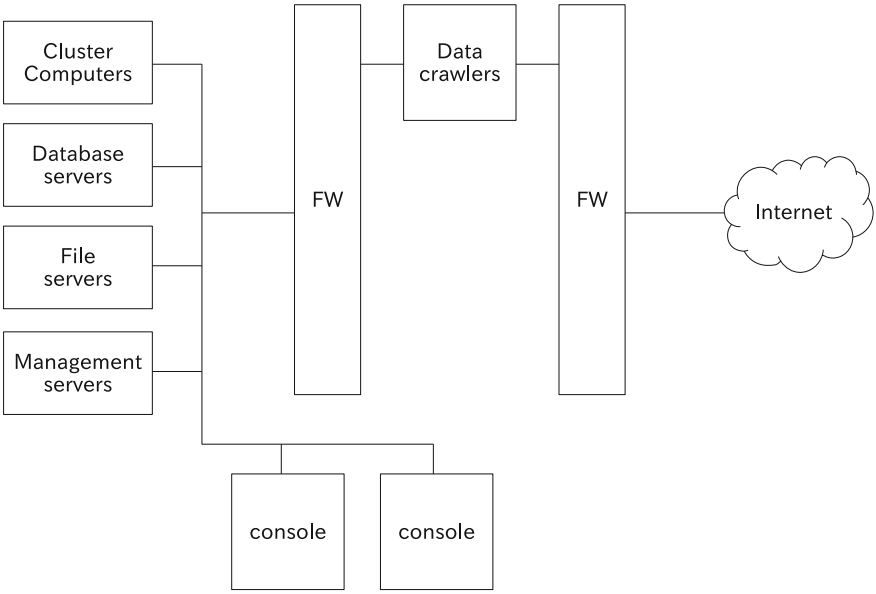


Fig. 4.2 A conceptual illustration of the research platform

4.1.1 Hardware

Several types of equipments are used to collect, accumulate, analyse data. A minimal research platform consists of two types of components: a database server and analysis software.

Figure 4.2 shows a conceptual illustration of an ideal research platform on data-centric study. This platform consists of data crawlers, cluster computers, database servers, management servers, file servers, and consoles.

- The data crawlers collect socioeconomic-technological data from e-commerce platforms, Web pages, blogs services, Web APIs or sensor networks in the Internet
- The raw data is stored as a file in the file server
- The data stored as files are cleaned and validated before being moved to the database server
- The data in the database server is analysed by analysis software
- Time-consuming computation (inference, estimation, and optimisation) is performed at the cluster computer in a parallel manner
- The management server controls these operations among the data crawlers, the file server, the database server and the cluster computer
- Researchers use this system from the consoles

The manufacturing and management activities of enterprises continue to rely on a much greater degree on high-quality data. In fact, high-quality data have become the basis which business is carried out, and benefit companies. The data-centric social sciences also provide methods to determine the quality of data.

Sensor networks represent another type of hardware that businesses rely on. Sensor networks involve both humans and machines. Human sensors are based on the concept of Goodchild [8]. In recent years, there has been an explosion of interest in using the Web to create, assemble, and disseminate geographical information provided voluntarily by individuals. He proposes a special term called volunteered geographic information (VGI), which is a special case of the more general Web phenomenon of user generated content. Machine sensors have been recently developed and used in our real lives; to provide examples, web cameras and web sensors continue to gain in popularity, and are more popular now than when they were first introduced. The use of these and other web applications can facilitate the collection of data on socioeconomic and environmental systems.

4.1.2 Software

In the typical application, the number of data points is large and data generation occurs at a speed faster than that possible with human computational skills. For these reasons, data should be handled and computed automatically. In order to improve the reliability of a data processing system, we should construct the system using several kinds of computer languages. For example, data should be stored in database servers and handled using an SQL. Statistical analysis of data should be conducted by a statistical computing language such as R. Finally, a script language should be used for analysis control and visualisation of results.

Given this circumstance, we need a technique of a (completely or partially automatic) batch process. The batch process can be divided into two types of processes (data processing and statistical analysis). Computer software has been specifically developed for certain disciplines and uses. Therefore, we should select computer software for their purposes. Roughly speaking, script languages (Perl, Ruby, PHP and

Python) should be used when we handle other software automatically, and statistical computation languages such as R, S-plus, Matlab and Octave should be used for the analysis of data. When we extract data from a database with filters, we can use languages to handle the database server. The SQLs are commonly implemented in relational database management systems (RDBMS). The Apache Hadoop software allows for the distributed processing of large data sets across clusters of computers using simple programming models. Several useful open software applications can be used for the purpose of data analysis and data processing. The following list contains examples of open source software to construct a system to realise both the data processing and data analysis.

- Perl; http://www.perl.org/
- Ruby; https://www.ruby-lang.org/
- PHP; http://www.php.net/
- Python; http://www.python.org/
- R; http://www.r-project.org/
- Octave; http://www.gnu.org/software/octave
- PostgreSQL; http://www.postgresql.org/
- MongoDB; http://www.mongodb.org
- MySQL; http://dev.mysql.com/
- Apache Hadoop; http://hadoop.apache.org/

In the next section, I will address each process: data acquisition, data collection and data analysis.

4.2 How to Acquire Data

At first, we need to either create (or obtain) data from the actual environment or acquire data from somewhere else. The data on socioeconomic-technological systems are collected from a multitude of locations:

- Open data sources (free of charge)
- Commercial data sources (that is, data are purchased)
- The actual environment, through the use of sensors (allowing generation of data)
- Crawling technology (collecting data from the Internet)

Some data sources provide an Web application programming interface (WebAPI) to obtain or purchase data as computer-readable files. Other data collection alternatives include downloading texts or crawling web pages for files. In this case, it is necessary to develop crawling software.

We have three types of methods to acquire data from data providers.

- as media provided from data providers
- via WebAPI
- as a file (CSV, TSV, Excel, XML, and so on) downloaded from a Web service

The situations for all socioeconomic-technological system sectors have changed or are currently changing from data-poor to data-rich. This permits data-centric social science research to be conducted based on large amounts of data related to these systems.

Recently, we have been able to purchase data from some data providers as CDs, DVDs, or USB sticks. We copy these files from the physical media onto our computers. When APIs provided by data providers are used, the additional need exists to develop computer programs to acquire the data from the servers of the data providers.

This allows us to calibrate model parameters from actual data. Moreover, the data integration from different kinds of data sources (such as demographic data and socioeconomic activity with geographic data) can be archived. Mesh statistics are detailed spatial statistics, which contain geographical information. These data allow us to detect statistical properties of each grid associated with the location. Grid statistics related to the census population, the population of different cities, the population of workers, economic values and transportation data can be used in this analysis. Map-Reduce architecture is one of implementations of parallel computing for large amounts of data. We further need to explore parallel computation combining database servers.

Normally, the computation of integrated data requires rich computational resources, including parallel computing techniques and a supercomputer.

4.3 Database Server and SQL

Many data elements cannot be handled as text files. I would like to recommend use of a database server to manage several types of datasets. Fundamental commands often used in a database server are as follows:

- create a table
- insert data into table
- search data
- update data
- delete data

4.3.1 Create a Table

When we use a database server, we need to first define the type of data. This process is called database normalisation. Normalisation usually involves dividing large tables into smaller tables and defining relationships between them. Codd [3] introduced a relational model of data and proposed the concept of normalisation at the same time [3]. Nowadays, this is realised as RDBMS.

In many relational database servers, this is done by defining a table, which is sometimes called schema. In the case of RDBMS, the data format is defined as a table consisting of columns and rows. If the data is stored in a tabular format (format

such as CSV, TSV and Excel), we can easily define the data format. In the case of XML format, we may use an XML database server or transform the XML data format into the relational data.

I will show several examples with concrete SQL commands in PostgreSQL, which is one of RDBMSs below. References [5, 12] provide a concise overview to help the readers understand and use PosegreSQL's features.

The SQL commands to create tables (weather and cities) are written as

```
CREATE TABLE weather (
  city varchar(80),
  temp_lo float,
  temp_hi float,
  prcp float,
  date date
);
CREATE TABLE cities (
  name varchar(80),
  latitude float,
  longitude float
);
```

In this example, the table `weather` contains five kinds of fields: `city`, `temp_lo`, `temp_hi`, `prcp`, and `date`. The table `cities` have three kinds of fields: `name`, `latitude` and `longitude`.

4.3.2 Insert Data into Table

The data is inserted into the table by using the 'INSERT' command:

```
INSERT INTO weather VALUES ('Kyoto',25.7,32.8,57,
'2013-08-01');
```

This is an implicit way but an explicit way is readable for programmers:

```
INSERT INTO weather (city,temp_lo,temp_hi,prcp,date)
VALUES ('Kyoto',25.7,32.8,57,'2013-08-01');
```

Table 4.1 PHP script to insert records coded as a CSV file into a table of the database server

```
#!/usr/bin/php -q
<?php
if(count($argv)!=2){
 print "$argv[0] datfile\n";
}
else{
 $file=$argv[1];
 if(file_exists($file)){
  $q = "host='localhost' port='5432' dbname='test'
        user='username' password='pass'";
  $h = pg_connect($q);
  if($h == false){
   print "connection error";
   exit;
  }
  $fp = @fopen($file,"r");
  $ff = preg_split("/\./",$file);
  $table = $ff[count($ff)-2];
  if($fp){
   while(!feof($fp)){
    $aline = trim(fgets($fp,30000));
    if(strlen($aline)>3){
     if(substr($aline,0,1)=='#'){
      $aline = substr($aline,1,strlen($aline)-1);
      $fieldname = preg_split("/[,;\t]/",$aline);
     }
     else{
      $a = preg_split("/[,;\t]/",$aline);
      $sql = sprintf("INSERT INTO %s (",$table);
      for($i=0;$i<count($fieldname);$i++){
       if($i==0) $sql = $sql . $fieldname[$i];
       else $sql = $sql . ", " . $fieldname[$i];
      }
      $sql = $sql . ") VALUES (";
       for($i=0;$i<count($a);$i++){
       $a[$i] = trim($a[$i]);
       $a[$i] = str_replace("\"","",$a[$i]);
       $a[$i] = str_replace("'","\"",$a[$i]);
       if($i==0) $sql = $sql . "'". $a[$i] . "'";
       else $sql = $sql . "," . "'" . $a[$i] . "'";
      }
      $sql = $sql . sprintf(");");
      print "$sql\n";
      $result = pg_exec($sql);
     }
    }
   }
   fclose($fp);
  }
  pg_close($h);
 }
}
?>
```

Table 4.2 Data recorded in a sample file

City	temp_lo	temp_hi	prcp	Date
Kyoto	25.7	32.8	57	2013-08-01
Kyoto	24.3	33.6	47	2013-08-02
Kyoto	23.2	33.3	44	2013-08-03
Kyoto	34.9	24.1	47	2013-08-04
Tokyo	26	32.8	63	2013-08-01
Tokyo	23	29.1	63	2013-08-02
Tokyo	23.5	30.8	53	2013-08-03
Tokyo	25.4	31.3	64	2013-08-04
Osaka	27.7	33.8	59	2013-08-01
Osaka	26.8	34.1	47	2013-08-02
Osaka	24.6	33.4	48	2013-08-03
Osaka	26	34.6	47	2013-08-04

Normally, to insert a large number of records automatically, we will call "**INSERT**" commands in a code described as a script language. Table 4.1 shows the PHP script to insert records coded as a CSV file into the database server. References [20, 24] contains a concrete information on PHP. In this PHP script, the dbname of the database server is assumed to be `test`, `localhost`, `username` and `pass` should be replaced in accordance with the server setting. The file name of the CSV file is assumed to be '`weather.csv`', which contains records shown in Table 4.2. The first line represents field names starting with #. The first command argument of the PHP script is assumed to be the file name of the CSV file.

4.3.3 Search Data

Searching data from a table in the database server is mostly used in data analysis. The "**SELECT**" command searches records that fulfil certain pre-selected conditions. Furthermore, selecting fields in a table is often used. In the following cases, we can search the data from weather and city.

```
SELECT * FROM weather WHERE temp_lo < 10;
SELECT city, temp_lo, temp_hi, date FROM weather;
```

Normally, we will call the "SELECT" command in codes written in R for data analysis. I will show an example code in Sect. 4.5.

4.3.4 Update Data

When we update the data in the table, we use the "**UPDATE**" command. In the case of subtracting two from the current values of `temp_hi` and `temp_lo` for data after 1999-12-3, we use the following SQL command.

```
UPDATE weather SET temp_hi = temp_hi - 2,
temp_lo = temp_lo -2 WHERE date > '1999-12-3';
```

We often use the "UPDATE" command to modify records when we find corrections of data after data validation and data verification.

4.3.5 Delete Data

If we want to delete data records, then use the "**DELETE**". For example, when we delete data records having `city='Tokyo'`, we use the following command:

```
DELETE FROM weather WHERE city = 'Tokyo';
```

We also delete all the elements from the table `weather` by using the following command:

```
DELETE FROM weather;
```

4.4 Analysis Software

There are several choices of analysis software as well as the database server. In this section, we will show several examples of codes with R. The statistical analysis software R is provided via R-Project.[1] The R for three different operating systems (Linux, Mac OS X, and Windows) are provided. After installation following the manual, R is executed by entering "R" from a command line. The R supports various kinds of libraries and there is dependency on the version of R. I show a part of libraries which are useful in data analysis. There are useful textbooks of R [1, 4] and documentations available at web pages of R-Project.

[1] http://r-project.org.

4.4.1 Packages

When we want to install an additional package named as "packagename" from the command line, we type:

```
> install.packages("packagename")
```

After installing the library "packagename", we load it using the following command:

```
> library(packagename)
```

The current version of R (version 3.0) has various libraries for data formats [6, 17, 18, 25], visualisation [7, 10, 21], parallel computation [16, 22], interfaces to other languages [14, 15, 19] and data analysis [2, 9, 11, 23]. The following list shows a part of the libraries in R:

- gdata: Various R programming tools for data manipulation [6]
- RMySQL: R interface to the MySQL database [17]
- RPostgreSQL: R interface to the PostgreSQL database system [18]
- XML: Tools for parsing and generating XML within R [25]
- Rmpi: An interface (wrapper) to MPI (Message-Passing Interface) [16]
- snow: Support for simple parallel computing in R [22]
- ggplot2: A plotting system for R [7]
- scatterplot3d: Plots a three dimensional (3D) point cloud [21]
- lattice: Data visualisation system for multivariate data [10]
- rPython: An interface (wrapper) to call Python from R [19]
- Rcpp: Seamless R and C++ Integration [14]
- Rgnuplot: R interface for gnuplot [15]
- bigmemory: Create, store, access and manipulate massive matrices [2]
- igraph: Network analysis and visualisation [9]
- maptools: Tools for reading and handling spatial objects [11]
- spdep: Spatial analysis for statistics and models [23]

4.4.2 Data Import

We need to import the text file into R before we start our working. The "**read.table**()" command is used for importing a text file into R. We will see an example of R commands. Assume that the text file is named topix.txt. The topix.txt is a text containing five cells inside the table which are separated by blank characters.

Table 4.3 Example of data on TOPIX

Date	Start_price	High_price	Low_price	End_price
2013/09/10	1181.31	1192.16	1180.74	1190.22
2013/09/11	1199.03	1199.3	1186.47	1189.25
2013/09/12	1187.35	1188.69	1178.87	1184.36
2013/09/13	1179.26	1190.27	1174.51	1185.28
2013/09/17	1190.55	1192.35	1181.56	1181.64
2013/09/18	1188.95	1201.08	1186.05	1193.07
2013/09/19	1205.67	1215.48	1199.63	1215.48
2013/09/20	1219.37	1221.8	1215.29	1218.98
2013/09/24	1208.67	1217.9	1206.47	1214.87
2013/09/25	1212.91	1214.66	1207.67	1211.15
2013/09/26	1200.92	1220.49	1191.63	1220.49
2013/09/27	1220.77	1223.12	1214.69	1217.52
2013/09/30	1200.15	1205.16	1192.28	1194.1
2013/10/01	1198.18	1204.32	1193.31	1193.44
2013/10/02	1193.99	1199.88	1171.06	1175.16
2013/10/03	1174.59	1180.26	1171.45	1173.99

Table 4.3 shows an example of data on TOPIX downloaded from Yahoo! Finance.[2] These contain dates and daily candles of TOPIX (start price, high price, low price and end price).

We can load the data into workspace with the "**read.table**()" command:

```
> data<-read.table("topix.txt", sep=" ", header=T)
```

The `sep` parameter specifies what character is used to separate columns in the file. The header parameter selects the existence of header characters. "**head**()" command is used for showing the table up to 5 rows.

```
> head(data, 5)
```

Table 4.4 shows an output of the head command.

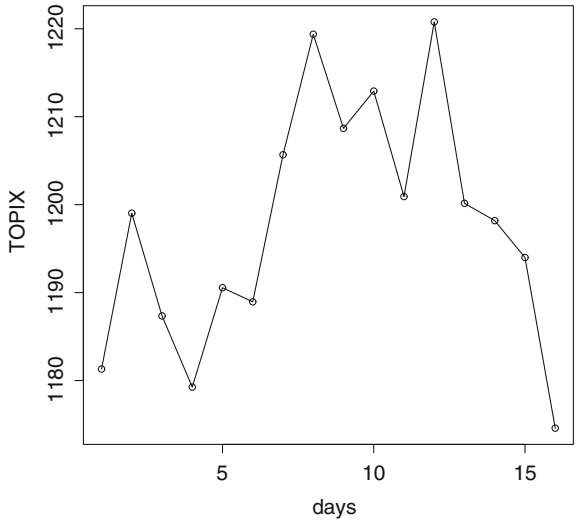
When you want to know details of the "**read.table**()" command, you can consult the R documentation.

[2] The time series data is downloaded from http://stocks.finance.yahoo.co.jp/stocks/history/?code=998405.T.

Table 4.4 Output of **head**() command in R

	Date	Start_price	High_price	Low_price	End_price
1	2013/09/10	1181.31	1192.16	1180.74	1190.22
2	2013/09/11	1199.03	1199.3	1186.47	1189.25
3	2013/09/12	1187.35	1188.69	1178.87	1184.36
4	2013/09/13	1179.26	1190.27	1174.51	1185.28
5	2013/09/17	1190.55	1192.35	1181.56	1181.64

Fig. 4.3 Example of a chart




```
> help(read.table)
```

4.4.3 Visualisation

R provides the "**plot**" command that is used to create time series charts. The "**plot**" command needs lists to serve as the x- and y-axes of the chart.

```
> plot(data[, 2],type="o",xlab="days",ylab="TOPIX")
```

Figure 4.3 shows a daily chart of start prices of TOPIX.

4.5 Examples

This section reviews the acquisition, collection and analysis of data on socioeconomic-technological systems. Two types of data analysis (flight-time tabular data and population data) are examined.

4.5.1 Data Analysis of a Flight Time Table

4.5.1.1 Data Acquisition

The flightaware provides a commercial Web API service called FlightXML.[3] Functions exist that enable us to obtain the flight time table in FlightXML. Using these functions of Web API allows collection of the flight time table data. We assume that we can collect the time table for flights and that this table will include 2,025 flights under the following conditions:

- Data process:
  1. Data on the flight time table were obtained from FlightXML of flightaware
  2. Data for all flights (both departing and arriving) at 58 airports in Japan were collected
  3. Airport-specific data on times of arrival and departure were collected
- Duration: 0:00–23:59 on 15 October, 2013 (UTC+9)
- Period of data collection: 5 hours

Table 4.5 shows a part of the data. The data contain identification code of a flight, actual identification code of a flight (for the code share flight), departure time, arrival time, a departure airport, an arrival airport, aircraft type, meal availability, the number of seats in first class, the number of seats in business class, and the number of seats in economy class.

4.5.1.2 Data Collection

We define the table named "`flightschedules`". Table 4.6 shows the definition of the table. After inserting all the data into the table "`flightschedules`" in the PostgreSQL database server, we processed and analyse the data. To count the daily number of connections between two airports for 58 Japanese airports, the "**SELECT**" command is used. Table 4.7 shows the SQL command and its result.

We need all the possible pairs of airports included in both the origin and destination fields. To extract the airport codes included in the data, we can use "**DISTINCT**"

[3] FlightAware: http://flightaware.com.

Table 4.5 An example of flight time table data

```
JAL6101;GK101;1381785900;1381791000;RJAA;RJBB;A320;;0;0;150
ANA8519;;1381836300;1381841700;RJAA;RJBB;B763;;0;0;0
NCA283;;1381837200;1381842300;RJAA;RJBB;B744;;0;0;0
DLH8382;DLH8383;1381841100;1381846500;RJAA;RJBB;MD11;;0;0;0
DLH8383;;1381841100;1381846500;RJAA;RJBB;MD11;;0;0;0
JAL6113;GK113;1381788600;1381794900;RJAA;RJCC;A320;;0;0;150
APW8521;;1381788900;1381795200;RJAA;RJCC;A320;;0;0;180
SKY871;;1381793700;1381799700;RJAA;RJCC;;;0;0;122
```

From *left* to *right* each field corresponds to identification code of a flight, actual identification code of a flight (for the code share flight), departure time, arrival time, a departure airport, an arrival airport, aircraft type, meal availability, the number of seats in first class, the number of seats in business class, and the number of seats in economy class

Table 4.6 The table definition for the flight time table

```
CREATE TABLE flightschedules (
      ident varchar(16),
      actual_ident varchar(16),
      departuretime integer not null,
      arrivaltime integer not null,
      origin varchar(6),
      destination varchar(6),
      air_crafttype varchar(8),
      seats_cabin_first integer,
      seats_cabin_business integer,
      seats_cabin_coach integer
);
```

Table 4.7 An example of an SQL command to count the number of connections from RJAA (Narita Airport) to RJBB (Kansai International Airport)

```
# SELECT COUNT(*) FROM flightschedules WHERE origin='RJAA'
and destination='RJBB';
count
-------
     5
(1 row)
```

in the "**SELECT**" command. Furthermore we can count the number of connections for all the possible combinations of airports as shown in Table 4.8. We obtained a list of 58 Japanese airports used in domestic flights.[4] Finally, 261 links among

[4] The included airports are listed as Narita International Airport (RJAA), Hyakuri Airport (RJAH), Kansai International Airport (RJBB), Nanki Shirahama Airport (RJBD), Kobe Airport (RJBE),

Table 4.8 An example of SQL command to detect all the airport codes included in the flight time table data and count the number of connections for each pair of connections

```
SELECT origin,destination,count(*) as cnt FROM
flightschedules WHERE origin IN (SELECT DISTINCT
origin FROM flightschedules) AND destination IN
(SELECT DISTINCT destination FROM flightschedules)
GROUP BY origin,destination ORDER by origin,destination;
```

Table 4.9 The daily number of connections

Departure airport	Arrival airport	# connections
RJAA	RJBB	5
RJAA	RJCC	15
RJAA	RJFF	15
RJAA	RJFO	2
RJAA	RJGG	15
RJAA	RJNK	6
RJAA	RJOA	6
RJAA	RJOM	2
RJAA	RJOO	15
RJAA	RJSN	3
RJAA	RJSS	11

Japanese domestic flights are extracted from the data. Table 4.9 shows a part of data on the daily number of connections. This can be represented as a weighted adjacency matrix.

(Footnote 4 continued)
Tokachi-Obihiro Airport (RJCB), New Chitose Airport (RJCC), Hakodate Airport (RJCH), Kushiro Airport (RJCK), Memanbetsu Airport (RJCM), Nakashibetsu Airport (RJCN), Wakkanai Airport (RJCW), Iki Airport (RJDB), Yamaguchi Ube Airport (RJDC), Tsushima Airport (RJDT), Asahikawa Airport (RJEC), Fukue Airport (RJFE), Fukuoka Airport (RJFF), Kagoshima Airport (RJFK), Miyazaki Airport (RJFM), Oita Airport (RJFO), Kitakyūshū Airport (RJFR), Saga Airport (RJFS), Kumamoto Airport (RJFT), Nagasaki Airport (RJFU), Chubu Centrair International Airport (RJGG), Nagoya Airport (RJNA), Komatsu Airport (RJNK), Shizuoka Airport (RJNS), Toyama Airport (RJNT), Noto Airport (RJNW), Hiroshima Airport (RJOA), Okayama Airport (RJOB), Izumo Airport (RJOC), Miho Yonago Airport (RJOH), Iwakuni Kintaikyo Airport (RJOI), Kchi Ryma Airport (RJOK), Matsuyama Airport (RJOM), Osaka International Airport (RJOO), Tottori Airport (RJOR), Tokushima Airport (RJOS), Takamatsu Airport (RJOT), Aomori Airport (RJSA), Yamagata Airport (RJSC), Fukushima Airport (RJSF), Hanamaki Airport (RJSI), Akita Airport (RJSK), Misawa Air Base (RJSM), Niigata Airport (RJSN), Odate Noshiro Airport (RJSR), Sendai Airport (RJSS), Shonai Airport (RJSY), Hachijojima Airport (RJTH), Tokyo International Airport (RJTT), Naha Airport (ROAH), Ishigaki Airport (ROIG), Kumejima Airport (ROKJ), and Miyako Airport (ROMY).

Table 4.10 Examples of functions to compute centrality measure of "**igraph**" in R

```
degree(g) # degree centrality
closeness(g) # closeness centrality
evcent(g)$vector # eigenvector centrality
page.rank(g, directed=TRUE)$vector # page rank
betweenness(g) # betweenness centrality
bonpow(g, exponent = 0.2) # Bonacich Power Centrality
alpha.centrality(g, alpha=1.0) # alpha centrality
```

4.5.1.3 Visualisation and Data Analysis

We focus on two types of visualisation to display the network structure:

1. Centrality
2. Weights

The centrality measures importance of nodes (See Sect. 3.3.4). The weights describe the property of links. "**igraph**" in R supports many types of centrality measures. Table 4.10 shows functions to compute centrality measures supported in "**igraph**".

A way in which centrality measures in Japanese domestic air transportation can be visualised will become apparent. Figure 4.4 shows the Japanese domestic air transportation network with the four different ways of visualisation: degree centrality, eigenvector centrality, alpha centrality and Page rank. The link weights are drawn in proportion to the daily number of connections between two airports. The size of nodes is in proportion to the centrality measure.

We can find that RJTT has the largest centrality in all the airports used in Japanese domestic air transportation for (a) degree centrality, (b) eigenvector centrality and (c) alpha centrality. In the case of alpha-centrality for $\alpha = 1$, RJTT does not have the largest value of centrality. The airports having the second and third largest centrality values depend on a type of centrality measure.

The histogram of the daily number of connections is shown in Figure 4.5. The maximum daily number of connections is 15. The small number of connections corresponds to flights to local airports. Depictions of larger numbers of connections correspond to flights between hub airports.

4.5.2 Data Analysis of Population

4.5.2.1 Data Acquisition

Next, we present an example in which population statistics for people living in 1-km^2 are used. These data are generated from the Population Census, conducted by the Statistics Bureau of Japan. The data are obtained from e-Stat (the portal site of government statistics) [13].

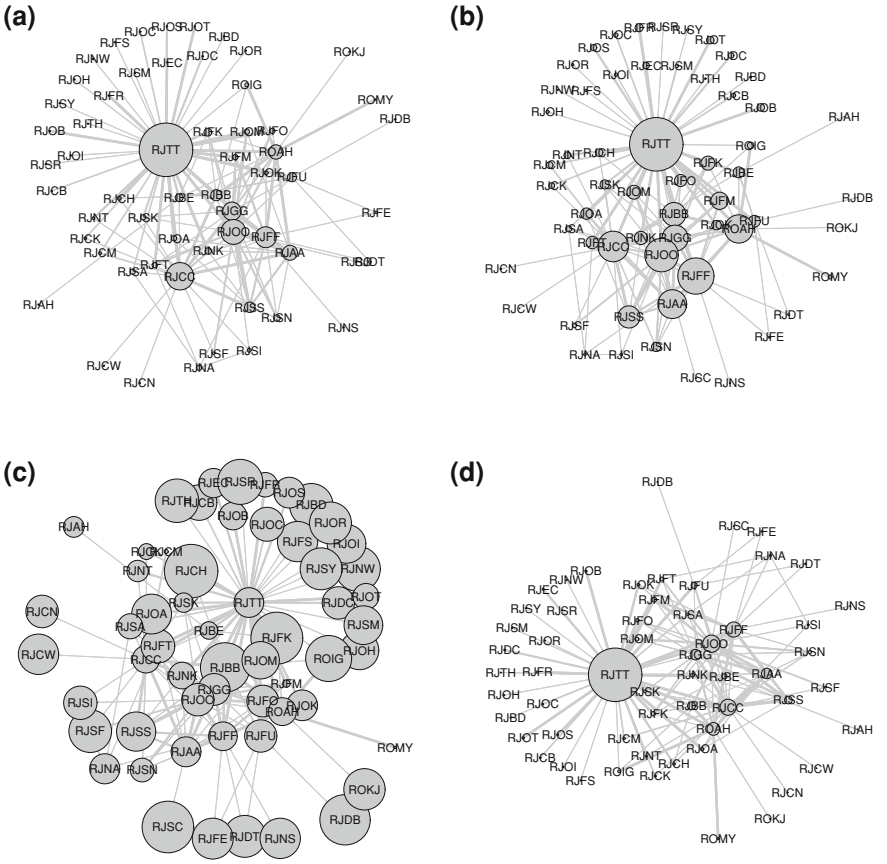


Fig. 4.4 The domestic air transportation network of Japan. The thickness of links is proportional to the daily number of flights. The size of nodes represents a value of, **a** degree centrality at the node, **b** its eigenvector centrality, **c** its alpha centrality ($\alpha = 1.0$) and **d** its Page Rank. The network consists of 58 nodes and is drawn from Japanese domestic flight data on 15 October, 2013

- Data process:
  1. Each record contains grid square code (corresponding to the longitude and latitude), the number of people living in the area of the grid square, the number of males, the number of females, and the number of families.
  2. In total, these data contain 180,220 records related to population, male population, female population, and the number of families.
  3. The data were collected for the year 2010, and are based on the Japanese population census for that year.
- Duration: 2010

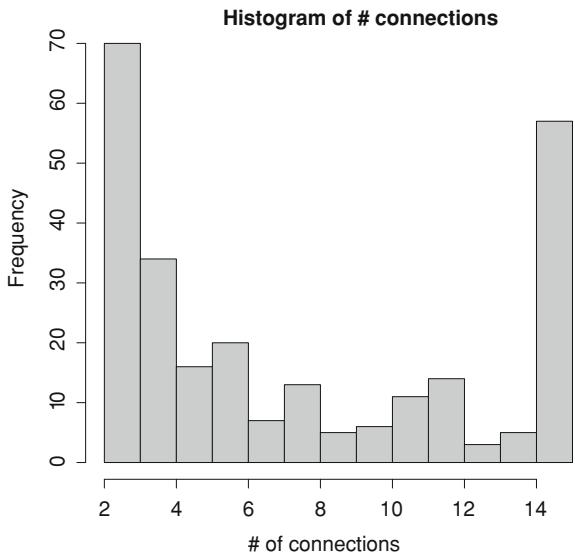


Fig. 4.5 The histogram of the daily number of connections extracted from the data in the flight time table for Japanese domestic flights

Table 4.11 The data schema definition for Japanese population (1-km^2 grid square data)

```
CREATE TABLE tblT000608 (
keycode varchar(10),
population int4,
male int4,
female int4,
family int4
);
```

4.5.2.2 Data Collection

In order to store our data to the PostgreSQL database server, we define the table "`tblT000608`" as shown in Table 4.11.

The first field "`keycode`" represents a grid square code; the second "`population`", the number of people living in the 1-km square represented by the "`keycode`"; "`male`", the number of males; "`female`", the number of females; and "`family`", the number of families. Throughout this example, we assume that the host name of the database server is "`localhost`", the user name is "`username`", and the password is "`pass`".

The grid square code defines the relationship between a grid square and geographical position which is numerically described as latitude and longitude.[5] The

[5] In 1973, the Administrative Management Agency (the current Ministry of Internal Affairs and Communications) announced Standard Grid Square and Grid Square Code Used for the Statistics

grid square code is calculated from the following equations:

$$\lfloor \text{latitude} \times 60 \div 40 \rfloor = p \quad (p \text{ is two digits.}), \tag{4.1}$$
$$a = (\text{latitude} \times 60 \div 40 - p) \times 40, \tag{4.2}$$
$$\lfloor a \div 5 \rfloor = q \quad (q \text{ is one digit.}), \tag{4.3}$$
$$b = (a \div 5 - q) \times 5, \tag{4.4}$$
$$\lfloor b \times 60 \div 30 \rfloor = r \quad (r \text{ is one digit.}), \tag{4.5}$$
$$c = (b \times 60 \div 30 - r) \times 30, \tag{4.6}$$
$$\lfloor \text{longitude} - 100 \rfloor = u \quad (u \text{ is two digits.}), \tag{4.7}$$
$$f = \text{longitude} - 100 - u, \tag{4.8}$$
$$\lfloor f \times 60 \div 7.5 \rfloor = v \quad (v \text{ is one digit.}), \tag{4.9}$$
$$g = (f \times 60 \div 7.5 - v) \times 7.5, \tag{4.10}$$
$$\lfloor g \times 60 \div 45 \rfloor = w \quad (w \text{ is one digit.}), \tag{4.11}$$
$$h = (g \times 60 \div 45 - w) \times 45, \tag{4.12}$$

where $\lfloor \cdot \rfloor$ represents the maximum integer less than $\cdot$. Consequently, from Eqs. (4.1), (4.3), (4.5), (4.7), (4.9) and (4.11), the grid square code is constructed from a sequence:

$$\text{grid square code} = puqvrw. \tag{4.13}$$

From Eqs. (4.2), (4.4) and (4.6), we obtain

$$\text{latitude} = p \times 40 \ \div 60 + q \times 5 \div 60 + r \times 30 \div 3600 + c \div 3600. \tag{4.14}$$

From Eqs. (4.8), (4.10) and (4.12), we get

$$\text{longitude} = 100 + u + v \times 7.5 \div 60 + w \times 45 \div 3600 + h \div 3600. \tag{4.15}$$

Therefore, in the case of statistics for the 1-km^2 area, since we have $0 \leq c < 30$ and $0 \leq h < 45$, we can decode both latitude and longitude at the bottom left corner of the grid from a grid square code by using the following equations:

$$\begin{aligned} \text{latitude}_0 = {} & (\text{1st and 2nd digits of grid square code}) \times 40 \ \div 60 \\ & + (\text{5th digit of grid square code}) \times 5 \div 60 \\ & + (\text{7th digit of grid square code}) \times 30 \div 3600, \end{aligned} \tag{4.16}$$

(Footnote 5 continued)
(Announcement No. 143 by the Administrative Management Agency on July, 12, 1973) as the integrated compilation method of the Grid Square. Also, this compilation method of the Grid Square was authorised as JIS in January, 1976 (Code JIS X 0410).

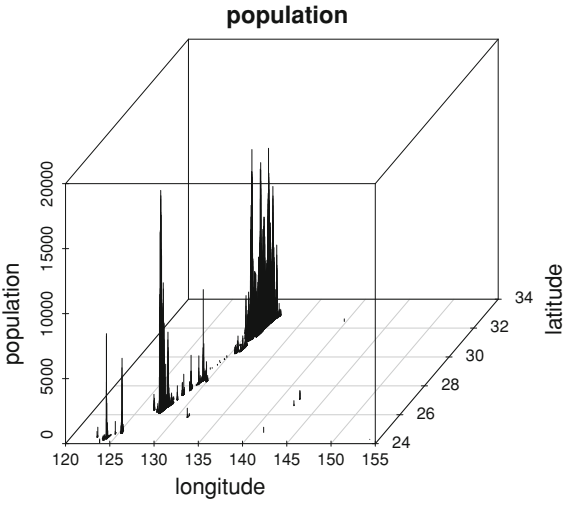


Fig. 4.6 The *x-axis* represents longitude, *y-axis* latitude, and *z-axis* population in the 1-km square. The displayed place corresponds to the southern part of Japan

$$\begin{aligned}\text{longitude}_0 = {} & 100 + (\text{3rd and 4th digits of grid square code}) \\ & + (\text{6th digit of grid square code}) \times 7.5 \div 60 \\ & + (\text{8th digit of grid square code}) \times 45 \div 3600. \end{aligned} \tag{4.17}$$

Thus, the ranges of latitude and longitude on the grid can be estimated as $\text{latitude}_0 \leq \text{latitude} < \text{latitude}_0 + 30/3600$ and $\text{longitude}_0 \leq \text{longitude} < \text{longitude}_0 + 45/3600$. This means that an interval of latitude is $30''$ and that an interval of longitude is $45''$.

4.5.2.3 Visualisation and Data Analysis

Visualisation is the first step in the understanding of data; we often visualise data in data validation and in explanatory data analysis.

To illustrate, let us attempt to draw a graph using 1-km^2 grid square area for the Japanese population as spatial data. Four types of data are available for this area: population, the number of males, the number of females, and the number of families. To draw this area, let us use the command "**scatterplot3d**()" from the library "**scatterplot3d**". This command generates a three-dimensional (3D) plot; we can use this command for 3D data.

We assume that all the records from this 1-km^2 grid square statistics of Japanese population census in 2010 was stored in our PostgreSQL database server. As mentioned in the above, the data is assumed to be inserted into the table "`tblT000608`" of dbname "`estat`" in a relational database server of localhost. Table 4.12 shows how to obtain the data from the PostgreSQL database server and compute Moran I. The "**RPostgreSQL**" library enables us to access a PostgreSQL database server. In this library, we mainly use the function "**dbConnect**()", "**dbSendQuery**()",

Table 4.12 R code for data visualisation

```
library(RPostgreSQL)
library(scatterplot3d)
# Obtain data from the database server
con <- dbConnect(PostgreSQL(),
                 host="localhost",
                 user="username",
                 password="pass",
                 dbname="estat")
dmy = dbSendQuery(con, "SET client_encoding = 'UTF-8'")
sql <- "SELECT keycode,population,male,
               female,family FROM tblT000608"
ans = dbSendQuery(con, sql)
mesh <- fetch(ans, n=-1)
dbDisconnect(con)
# Select place
N <- 20000
s <- 1
e <- N-s+1
x <- matrix(0,N,2)
z <- matrix(0,N,1)
# Translate latitude and longitude from grid square code
for(i in 1:N){
  c1 = mesh[i+s-1,]$keycode
#Decode longitude from a grid square code
  x[i,1] = 100+as.numeric(substr(c1,3,4))
  x[i,1] = x[i,1] + as.numeric(substr(c1,6,6))*7.5/60
  x[i,1] = x[i,1] + as.numeric(substr(c1,8,8))*45/3600
#Decode latitude from a grid square code
  x[i,2] = as.numeric(substr(c1,1,2))*40/60
  x[i,2] = x[i,2] + as.numeric(substr(c1,5,5))*5/60
  x[i,2] = x[i,2] + as.numeric(substr(c1,7,7))*30/3600
}
# plot
scatterplot3d(x[,1],x[,2],mesh$population[s:e],
main="population", xlab="longitude", ylab="latitude",
zlab="population", cex.lab=1.6, cex.axis=1.2, cex.main=1.8,
highlight.3d=F, type="h", pch = " ", lwd=1)
dev.copy2eps(file="population.eps")
dev.off()
```

"**fetch**()" and "**dbDisconnect**()". "**dbConnect**()" is used first to connect to the database server. "**dbDisonnect**()" is used to finally close the connection. "**dbSendQuery**()" is a function that sends a query to the database server with which the connection has been established in order to obtain query results.

Figure 4.6 shows the spatial distribution for the Japanese population in the southern part of Japan. This graph is generated by the following R code.

Next, let us compute Moran's I and Geary's C using an alternative method (See Sect. 3.3.5). We will use **moran.test**() and **geary.test**() in the library "**spdep**".

Table 4.13 R code for computing spatial autocorrelation based on Moran's I and Geary's C

```
library(RPostgreSQL)
library(spdep)
# Obtain data from DB
con <- dbConnect(PostgreSQL(),
                 host="localhost",
                 user="user",
                 password="passwd",
                 dbname="estat")
dmy = dbSendQuery(con, "SET client_encoding = 'UTF-8'")
sql <- "SELECT keycode,population,male,
               female,family FROM tblT000608"
ans = dbSendQuery(con, sql)
mesh <- fetch(ans, n=-1)
dbDisconnect(con)
# Select place
N <- 20000
s <- 1
e <- N-s+1
x <- matrix(0,N,2)
z <- matrix(0,N,1)
# Translate latitude and longitude from grid square codes
for(i in 1:N){
  c1 = mesh[i+s-1,]$keycode
#Decode longitude from a grid square code
  x[i,1] = 100+as.numeric(substr(c1,3,4))
  x[i,1] = x[i,1] + as.numeric(substr(c1,6,6))*7.5/60
  x[i,1] = x[i,1] + as.numeric(substr(c1,8,8))*45/3600
#Decode latitude from a grid square code
  x[i,2] = as.numeric(substr(c1,1,2))*40/60
  x[i,2] = x[i,2] + as.numeric(substr(c1,5,5))*5/60
  x[i,2] = x[i,2] + as.numeric(substr(c1,7,7))*30/3600
}
# Obtain Doronei triangle
lp.tri.nb <- tri2nb(x)
# Display Moran's I
m<-moran.test(mesh$population[s:e],
    nb2listw(lp.tri.nb,style="W"))
cat(sprintf("moran population %f %f %f %f\n",
    m$estimate[1],m$estimate[2],m$estimate[3],m$p.value))
# Display Geary's C
g<-geary.test(mesh$population[s:e],
    nb2listw(lp.tri.nb,style="W"))
cat(sprintf("geary population %f %f %f %f\n",
    g$estimate[1],g$estimate[2],g$estimate[3],g$p.value))
```

moran.test() and **geary.test**() do not need a geodesic distance matrix for all the features. In the place of this matrix, **moran.test**() and **geary.test**() use a neighbours list that is constructed by the command **tri2nb**(). This function constructs a *doronei* list to express the distance. This contributes to the reduction in the number of arrays in the computation. Table 4.13 shows a sample R code to compute both Moran I and Geary C. Table 4.14 shows results computed by the sample R code. We can see that the value of Moran's I is close to 1 and that the value of Geary's C is close to 0. This means that the features of the population are strongly concentrated with statistical significance since these p-values are zero.

Table 4.14 The values of spatial autocorrelation for Moran's I and Geary's C

Type	Moran's I	E[I]	$\sqrt{\text{Var}[I]}$	p-value
Population	0.729504	−0.000050	0.000017	0.000000
Type	Geary's C	E[C]	$\sqrt{\text{Var}[C]}$	p-value
Population	0.272013	1.000000	0.000044	0.000000

These values are obtained from **moran.test**() and **geary.test**() in "**spdep**" library of R. The data of 1-km^2 grid square of the Japanese population in 2010 is used

References

1. Adler, J.: R in a Nutshell: A Desktop Quick Reference. O'Reilly, Sebastopol (2010)
2. bigmemory: http://cran.r-project.org/web/packages/bigmemory/bigmemory.pdf. Accessed 15 Jan 2014
3. Codd, E.F.: A relational model of data for large shared data banks. Commun. ACM **13**, 377–387 (1970)
4. Cotton, R.: Learning R- A step-by-step Function Guide to Data Analysis. O'Reilly, Sebastopol (2013)
5. Drake, J.D., Worsley, J.C.: Practical PostgreSQL. O'Reilly, Sebastopol (2002)
6. gdata: http://cran.r-project.org/web/packages/gdata/gdata.pdf. Accessed 15 Jan 2014
7. ggplot2: http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. Accessed 15 Jan 2014
8. Goodchild, M. F.: Citizens as voluntary sensors- spatial data infrastructure in the world of Web 2.0. Int. J. Spat. Data Infrastruct. Res. 2, 24–32 (2007).
9. igraph: http://cran.r-project.org/web/packages/igraph/igraph.pdf. Accessed 15 Jan 2014
10. lattice: http://cran.r-project.org/web/packages/lattice/lattice.pdf. Accessed 15 Jan 2014
11. maptools: http://cran.r-project.org/web/packages/maptools/maptools.pdf. Accessed 15 Jan 2014
12. Obe, R., Hsu, L.: PostgreSQL- Up and Running. O'Reilly, Sebastopol (2012)
13. Portal Site of Official Statistics of Japan: http://www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do. Accessed 3 Mar 2014
14. Rcpp: http://cran.r-project.org/web/packages/Rcpp/Rcpp.pdf. Accessed 15 Jan 2014
15. Rgnuplot: http://cran.r-project.org/web/packages/Rgnuplot/Rgnuplot.pdf. Accessed 15 Jan 2014
16. Rmpi: http://www.stats.uwo.ca/faculty/yu/Rmpi/. Accessed 15 Jan 2014
17. RMySQL: http://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf. Accessed 15 Jan 2014
18. RPostgreSQL: http://cran.r-project.org/web/packages/RPostgreSQL/RPostgreSQL.pdf. Accessed 15 Jan 2014
19. rPython: http://cran.r-project.org/web/packages/rPython/rPython.pdf. Accessed 15 Jan 2014
20. Sanders, W.: Learning PHP Design Patterns. O'Reilly, Sebastopol (2013)
21. scatterplot3d: http://cran.r-project.org/web/packages/scatterplot3d/scatterplot3d.pdf. Accessed 15 Jan 2014
22. snow: http://cran.r-project.org/web/packages/snow/snow.pdf. Accessed 15 Jan 2014
23. spdep: http://cran.r-project.org/web/packages/spdep/spdep.pdf. Accessed 15 Jan 2014
24. Tatroe, K., Macintyre, P., Lerdorf, R.: Programming PHP, 3rd edn. O'Reilly, Sebastopol (2013)
25. XML: http://cran.r-project.org/web/packages/XML/XML.pdf. Accessed 15 Jan 2014

Part III
Exemplar Studies

Chapter 5
Risk Assessment of Extreme Events

Abstract Risk assessment is one of the crucial issues in management science. Specifically, it is important to infer risks of extreme events, which generate huge damage with small probability. To estimate risk of these extreme events, we need a method to extrapolate tail probabilities. In this chapter, the method to estimate parameters and empirical evidence are introduced through exemplar study of the foreign exchange market.

5.1 Introduction

There are more than 100 kinds of currencies in the world, and exchange rates between most of them are determined by trades through the market. Table 5.1 shows foreign exchange turnover in April 2013, which is provided by BIS Triennial Central Bank Survey [3]. Trading in foreign exchange markets averaged 5.3 trillion USD per day in April 2013. This is up from 4.0 trillion USD in April 2010 and 3.3 trillion in April 2007.

Usually, traders participate in the foreign exchange market all around the world, and they actively trade in their business hours. Therefore, there is a typical 24-h pattern in the trading activity. Since trades and quotes are mainly conducted in electronic systems, their exchange rates are changing second by second.

Basically, their fluctuations are based on the balance of long-term supply and demand, but a variety of factors seems to affect the exchange rates. Recently, it has been much easier for individuals to buy and sell currencies in the market, and it becomes more important to understand the foreign exchange risk to hold several types of currencies safely. Exchange rates sometimes fluctuate unpredictably, causing loss of value in holding currencies. Especially, it is well-known that the volatilities of the price fluctuate depending on the time period, which is observed as fat-tailedness of a probability density function (PDF) for log-return time series. Therefore, it is important to regard its fat-tailedness when we estimate the risk from historical data.

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_5, © Springer Japan 2014

Table 5.1 The turnover is adjusted for local and cross-border inter-dealer double-counting

Instrument	1998	2001	2004	2007	2010	2013
Foreign exchange instruments	1,527	1,239	1,934	3,324	3,971	5,345
Spot transactions	568	386	631	1,005	1,488	2,046
Outright forwards	128	130	209	362	475	680
Foreign exchange swaps	734	656	954	1,714	1,759	2,228
Currency swaps	10	7	21	31	43	54
Options and other products	87	60	119	212	207	307
Turnover at April 2013 exchange rates	1,718	1,500	2,036	3,376	3,969	5,345

The numbers express daily averages in April, in billions of US dollars

Foreign exchange rates have been investigated by numerous researchers with various approaches based on statistics and time series analysis. Mandelbrot and Taylor [14] proposed the concept of time changes or subordinated process in order to explain fat-tailedness of returns. According to the normal mixture model in finance, unconditional distributions of log-returns are reported to be well-fitted to the mixture of normal distributions with an unconditional distribution of volatility [5]. Beck proposed the same theory in the literature of superstatistics [2]. Gabaix et al. examine the power law distributions observed in financial markets [10] and proposes a model providing an explanation for these empirical power laws [11].

Tsallis statistics or nonextensive statistical mechanics is also useful to fit the unconditional PDF of log-return time series [12]. Drożdż showed that exchange rate return fluctuations for many currency pairs are well-described by nonextensive statistics [7]. In this framework, the q-Gaussian, a generalised normal distribution with index q, is used to explain the fat-tailedness of actual data. Here, we mainly focus on the q-Gaussian distribution for the purpose of risk estimation (see Sect. 5.3).

Suppose that we hold a certain foreign currency and that we can accept loss to a certain amount of money, as we possess some money used as risk buffer. Here, we introduce a method to compare and evaluate foreign exchange risk by using a loss probability where loss exceeds the risk buffer. Let $r(s) = \ln R(s+1) - \ln R(s)$ $(s = 0, \ldots, T-1)$ be a log-return at time s, where s represents a daily business day and $R(s)$ an exchange rate between pairwise currencies at time s.

Denoting $p(r)$ as an unconditional PDF of r, we can express the probability where incidental or short-term loss becomes larger than the buffer h as

$$\Pr[r < -h] = \int_{-\infty}^{-h} p(r')\mathrm{d}r'. \tag{5.1}$$

Using a definition of relative frequency for historical data, this can be approximated as

$$\Pr[r < -h] \approx \frac{N[r < -h]}{T}, \tag{5.2}$$

where $N[r < -h]$ is the number of losses that exceed the deposit h in the past, and T is the total number of historical data. In this empirical method, however, the probability for loss to be greater than the maximum loss in the historical data is always estimated as 0. Namely, Eq. (5.2) shows a truncation at the maximum loss. Because the number of historical data is finite, we cannot solve this truncation.

To assess the probability of loss larger than the maximum loss in the historical data, we need to use an extrapolation method for the ruin probability under a proper model assumption that can grasp fluctuations of the exchange rates. As the simplest model, $p(r)$ is often approximated as a Gaussian distribution with mean μ and standard deviation σ. In this case, Eq. (5.1) is described as

$$\Pr[r < -h] = \frac{1}{2}\mathrm{erfc}\Big(\frac{h-\mu}{\sqrt{2}\sigma}\Big), \tag{5.3}$$

where $\mathrm{erfc}(x)$ represents the complementary error function defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\mathrm{d}t.$$

Therefore, we can estimate the ruin probability from T historical observations $\{r(0), \ldots, r(T-1)\}$ by plugging sample mean and standard deviation into μ and σ of Eq. (5.3), respectively. These are estimated as

$$\hat{\mu} = \frac{1}{T}\sum_{s=0}^{T-1} r(s), \quad \hat{\sigma}^2 = \frac{1}{T}\sum_{s=0}^{T-1}\Big(r(s)-\hat{\mu}\Big)^2. \tag{5.4}$$

However, this model does not fit the empirical probability distribution computed from actual data of the foreign exchange rates. As an alternative model, we introduce the q-Gaussian distribution$(1 < q < 3)$, defined as

$$p_q(r; q, \mu_q, \sigma_q) = \frac{1}{B\Big(\frac{1}{q-1}-\frac{1}{2}, \frac{1}{2}\Big)}\sqrt{\frac{q-1}{(3-q)\sigma_q^2}}\Big(1+\frac{q-1}{(3-q)\sigma_q{}^2}(r-\mu_q)^2\Big)^{\frac{1}{1-q}}, \tag{5.5}$$

where $B(a, b)$ is the beta function[1] defined as

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t. \tag{5.6}$$

[1] The beta function is symmetric: $B(a, b) = B(b, a)$.

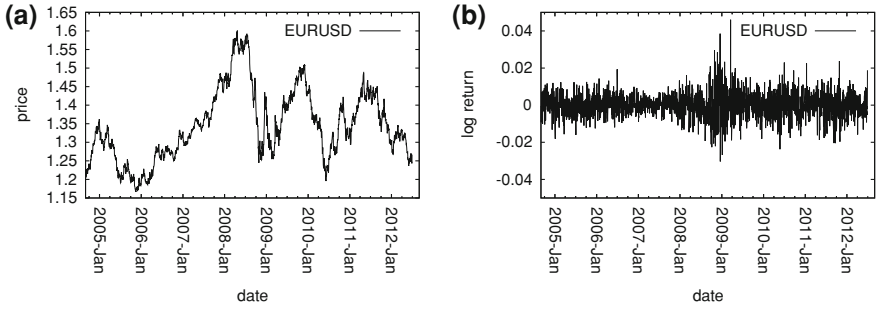


Fig. 5.1 **a** Daily exchange rates of EUR/USD for the period from September 2004 to August 2012 and **b** their daily log-returns

The q-Gaussian distribution for $1 < q < 3$ is equivalent to Student's t-distribution

$$p(x) = \frac{1}{\nu^{1/2} B(\frac{\nu}{2}, \frac{1}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \tag{5.7}$$

where $\nu = (3 - q)/(q - 1)$ is satisfied.

Before considering an unconditional distribution of log-return, we see a time series of the foreign exchange rates. Figure 5.1 shows the daily exchange rates of EUR/USD and their log-returns for the period of September 2004 to August 2012. Figure 5.2 shows its complementary cumulative distribution. This is calculated in three manners explained above (empirical distribution, Gaussian distribution and q-Gaussian distribution). We see that the Gaussian distribution does not fit the empirical distribution at all. This is because the Gaussian distribution ignores the influence volatility fluctuation observed in the exchange markets. As a result, it can underestimate the loss probability.

On the other hand, the q-Gaussian is well-fitted to the empirical distribution, and it can extrapolate the tail probability. From this extrapolation, we can assess the loss probability. In fact, the q-Gaussian is obtained by a mixture of normal distributions of which deviation σ fluctuates for a long time scale. This seems to be related to the fact that volatilities of the return observed from data in the market are not constant and depend on a time period. This may be the reason why the q-Gaussian is a good model for estimating foreign currency risk.

In this chapter, we discuss a parametric risk assessment procedure with the q-Gaussian and another efficient distribution, the Pearson type IV, and examine the foreign exchange risk for 30 currency pairs traded in the foreign exchange market.

In Sect. 5.3, we describe the framework of Tsallis statistics, which we use for risk estimation. In Sect. 5.4, we explain maximum likelihood procedure, the method to estimate parameters of the assumed distribution from given observations. In Sect. 5.6, we perform analysis of daily data for 30 currency pairs using the q-Gaussian distribution. In Sect. 5.7, we deal with the Pearson type IV and compare it with the q-Gaussian. Section 5.8 is devoted to the conclusion.

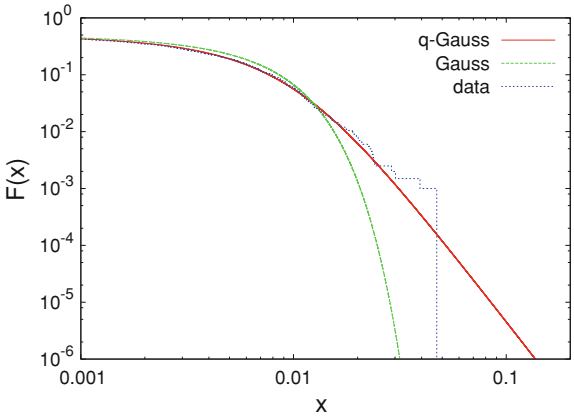


Fig. 5.2 Cumulative distributions of daily log-returns of EUR/USD for the period from September 2004 to August 2012 ($T = 2008$) calculated in three manners: empirical, Gaussian, and q-Gaussian distributions. A *solid curve* represents a fit by means of a q-Gaussian distribution. A *dashed curve* shows a Gaussian fitting, and a *dotted curve* the empirical

5.2 GARCH Processes

A generalised autoregressive conditional heteroskedastic (GARCH) process, which is a generalisation of an autoregressive conditional heteroskedastic (ARCH) process proposed by Engle [8], has been formalised by Bollerslev [4]. The GARCH(p,q) process is described as

$$\begin{cases} \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \\ x_t \ = \sigma_t z_t \end{cases}, \tag{5.8}$$

where x_t is a dynamical variable at discrete time t, σ_t is called volatility at time t, and z_t is an *i.i.d.* zero-mean standard normal random variable. In order to guarantee the positivity of unconditional mean of σ_t^2,

$$\mathrm{E}[\sigma^2] = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}, \tag{5.9}$$

parameters $\{\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\}$ should satisfy

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1. \tag{5.10}$$

The ARCH (p) process is also expressed as the GARCH (p, 0) process, which is described as

$$\begin{cases} \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 \\ x_t = \sigma_t z_t \end{cases}. \tag{5.11}$$

The simplest case of the ARCH process ($p = 1$) can be rewritten as

$$x_t = \sqrt{\alpha_0 + \alpha_1 x_{t-1}^2} \cdot z_t. \tag{5.12}$$

According to Haan [13], it is proven that the unconditional PDF of Eq. (5.12) has the power law tails,

$$p(x) \propto x^{-a-1}, \tag{5.13}$$

where a represents the power law exponent $a > 0$, which follows

$$\frac{\alpha_1^{a/2} 2^{a/2}}{\sqrt{\pi}} \Gamma\left(\frac{a+1}{2}\right) = 1. \tag{5.14}$$

The unconditional PDF in Eq. (5.13) is present for

$$\mathrm{E}[\ln(\sqrt{\alpha_1}|z|)] < 0. \tag{5.15}$$

In this derivation, the conditions for the power law exponent of Kesten random multiplicative processes [6, 21] is used.

5.3 Tsallis Statistics

In this section, we introduce a q-Gaussian distribution. Tsallis statistics is a generalisation of ordinary Boltzmann-Gibbs statistical mechanics to describe statistical behaviours of complex systems. Here, we introduce a q-Gaussian distribution, one of the typical distributions of generalised canonical distribution.

Tsallis [12] introduced a q-extension of exponential function defined as

$$\exp_q(x) \equiv (1 + (1-q)\, x)^{\frac{1}{1-q}}. \tag{5.16}$$

This is the solution of the ordinary differential equation, $\frac{\mathrm{d}y}{\mathrm{d}x} = y^q$, $y(0) = 1$. Then, $\ln_q x$ is defined as the inverse function of Eq. (5.16),

$$\ln_q x \equiv \frac{x^{1-q}-1}{1-q}. \tag{5.17}$$

Therefore, q–entropy, a generalised entropy functional, is defined as

$$S_q[p] \equiv \int_{-\infty}^{\infty} p(x)\ln_q\left(\frac{1}{p(x)}\right)\mathrm{d}x = \frac{1-\int_{-\infty}^{\infty} p(x)^q \mathrm{d}x}{q-1}. \tag{5.18}$$

In the limit of $q \to 1$, this functional converges into the Boltzmann-Gibbs entropy. From the maximisation of Eq. (5.18) under some constraints, we have the q-Gaussian distribution:

$$p(x) = \frac{1}{B\left(\frac{q}{q-1}-\frac{1}{2},\frac{1}{2}\right)}\sqrt{\frac{q-1}{(3-q)\sigma_q^2}}\left(1+\frac{q-1}{(3-q)\sigma_q{}^2}(x-\mu_q)^2\right)^{\frac{1}{1-q}}. \tag{5.19}$$

Equation (5.19) can also be described as

$$p(x) = A_q\sqrt{B_q}\exp_q\left(-B_q(x-\mu_q)^2\right), \tag{5.20}$$
$$A_q = \frac{\sqrt{q-1}}{B\left(\frac{q}{q-1}-\frac{1}{2},\frac{1}{2}\right)},\ B_q = \frac{1}{(3-q)\sigma_q^2}.$$

The derivation of the q-Gaussian distribution is shown in Appendix A.

As shown in Appendix B, the complementary cumulative distribution of the q-Gaussian for $1 < q < 3$, $F_1(x) = \Pr[X \geq x]$ is obtained as

$$F_1(x) = \begin{cases} \dfrac{1}{2}\left(1-\beta\left(\dfrac{\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2}{1+\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2};\dfrac{1}{2},\dfrac{1}{q-1}-\dfrac{1}{2}\right)\right) & (x \geq \mu_q) \\ \dfrac{1}{2}\left(1+\beta\left(\dfrac{\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2}{1+\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2};\dfrac{1}{2},\dfrac{1}{q-1}-\dfrac{1}{2}\right)\right) & (x < \mu_q), \end{cases} \tag{5.21}$$

where $\beta(x;a,b)$ is the regularised incomplete beta function defined as

$$\beta(x;a,b) = \frac{1}{B(a,b)}\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t \qquad (0 \leq x \leq 1).$$

Note that the q-Gaussian distribution is derived as a stationary distribution of the following Langevin equation [9]

$$\mathrm{d}u = -\nu u \mathrm{d}t + \sqrt{\gamma u^2 + \delta}\mathrm{d}W \ \ (\nu, \gamma, \delta > 0)\,, \tag{5.22}$$

where u is a dynamical variable and $W(t)$ is the Wiener process such that

$$\mathrm{E}[W(t)] = 0, \quad \mathrm{E}[W(t_1)W(t_2)] = \min(t_1, t_2). \tag{5.23}$$

Let us derive that its stationary distribution $p(u)$ obeys the q-Gaussian. The corresponding Fokker–Planck equation of Eq. (5.22) is described as

$$\frac{\partial p(u,t)}{\partial t} = \nu \frac{\partial}{\partial u}\Big[up(u,t)\Big] + \frac{1}{2}\frac{\partial^2}{\partial u^2}\Big[(\gamma u^2 + \delta)p(u,t)\Big]. \tag{5.24}$$

Then, the stationary distribution $p(u)$ is the solution of the following equation:

$$\nu \frac{\mathrm{d}}{\mathrm{d}u}\Big[up(u)\Big] + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\Big[(\gamma u^2 + \delta)p(u)\Big] = 0. \tag{5.25}$$

Under the natural boundary condition $p(\pm\infty) = 0$ and $\frac{\mathrm{d}p}{\mathrm{d}u}|_{u=\pm\infty} = 0$, this is naturally integrated as

$$\nu u p(u) + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}u}\Big[(\gamma u^2 + \delta)p(u)\Big] = 0,$$

which leads to

$$\frac{\mathrm{d}p(u)}{p(u)} = -\frac{2(\nu + \gamma)u}{\gamma u^2 + \delta}\mathrm{d}u.$$

Therefore, we have the solution

$$p(x) = C\left(1 + \frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}}, \tag{5.26}$$

where C is a normalisation constant. Equation (5.26) is the form of q-Gaussian distribution with $q = \frac{\nu+2\gamma}{\nu+\gamma}$ and $\sigma_q^2 = \frac{\delta}{2\nu+\gamma}$. This is a model that its deviation fluctuates with u. In the case of $\gamma = 0$ in Eq. (5.22), its stationary distribution is the Gaussian. This shows that fluctuation of the deviation results in fat-tailedness of the distribution.

5.4 Maximum Likelihood Method

In this section, a maximum likelihood method is introduced. This is a parameter estimation procedure from observations under an assumed distribution. We verify whether q-Gaussian distributions plugged parameter estimates are well-fitted to the log-return time series of foreign exchange rates.

Suppose that we estimate parameters (q, μ_q, σ_q) of the q-Gaussian distribution $p_q(x; q, \mu_q, \sigma_q)$ from T observations $\{r(0), \ldots, r(T-1)\}$. Here, we assume that $r(s)$ is sampled in *i.i.d.* Thus, the log-likelihood function is set as

$$l(q, \mu_q, \sigma_q) = \sum_{s=0}^{T-1} \ln p_q(r(s); q, \mu_q, \sigma_q). \tag{5.27}$$

By maximising $l(q, \mu_q, \sigma_q)$, we obtain parameter estimates of assumed PDF,

$$\{\hat{q}, \hat{\mu_q}, \hat{\sigma_q}\} = \arg \max_{q, \mu_q, \sigma_q} l(q, \mu_q, \sigma_q). \tag{5.28}$$

This solution can be obtained from the likelihood equations

$$\frac{\partial l}{\partial q} = \frac{\partial l}{\partial \mu_q} = \frac{\partial l}{\partial \sigma_q} = 0. \tag{5.29}$$

However, as we cannot get the solution in an analytical manner, we solve it numerically. In order to solve this optimisation problem, we use a gradient method as shown in Sect. 3.1.7.

Furthermore, since the log-likelihood function is multimodal, its convexity is not guaranteed. Therefore, we calculate optimised parameters from different initial parameter values in 50 trials, and then we choose the most optimal ones as parameter estimates.

5.5 Test with Artificial Data

Before we treat historical data in the market, we want to check the validity of applying this method. Here we examine whether we can estimate the parameters that we set with artificial data. We make datasets drawn from a q-Gaussian distribution with the Generalised Box–Müller method (GBMM) [22], in which samples are generated as

$$\begin{cases} Z_1 = \sqrt{-2\ln_{q'}(\zeta_1)}\cos(2\pi\zeta_2) \\ Z_2 = \sqrt{-2\ln_{q'}(\zeta_1)}\sin(2\pi\zeta_2) \end{cases}, \tag{5.30}$$

Table 5.2 Parameter estimates for datasets of samples drawn from the q-Gaussian by using GBMM, in which the parameter is set as $(q, \mu_q, \sigma_q) = (1.5, 2.0, 3.0)$

Size	q	μ_q	σ_q
10^2	1.430907	2.063977	3.068865
10^3	1.488220	2.006066	3.007677
10^4	1.499276	2.003754	2.999074
10^5	1.499567	2.000199	2.999297

where ζ_1 and ζ_2 are $i.i.d.$ $(0, 1)$ uniform random numbers, and $q = \frac{3q'-1}{q'+1}$. Then, Z_1 and Z_2 obey the q-Gaussian distribution with $\mu_q = 0$ and $\sigma_q = 1$, which we denote as $Z_1, Z_2 \sim N_q(0, 1)$. Then, $\sigma_q Z_1 + \mu_q \sim N_q(\mu_q, \sigma_q)$. Thus, we can generate q-Gaussian random numbers with any possible parameters.

Table 5.2 shows averages of the parameter estimates for the artificial q-Gaussian datasets, in which the parameters are set as $(q, \mu_q, \sigma_q) = (1.5, 2.0, 3.0)$. The data size is 10^2, 10^3, 10^4 and 10^5, and we estimate parameters for 100 datasets of each data size. The table indicates that we can estimate the parameters we set regardless of the data size.

5.6 Application of the q-Gaussian for the Foreign Exchange Market

We perform empirical data analysis in order to check the adequacy of applying q-Gaussian for foreign exchange markets. We analyse log-returns of daily closing price for the period from September 2004 to August 2012 of 30 currency pairs consisting of 11 currencies (see Table 5.3).[2]

The data are downloaded from PACIFIC Exchange Rate Service [16]. Using the method in Sect. 5.4, we estimate parameters of q-Gaussian distributions for the log-returns of the 30 pairs. Table 5.4 shows the results of parameter estimation.

We see that in all cases, parameter estimates $\hat{q}$ are larger than 1.3. This means that fluctuations of volatility cannot be ignored, and that the assumption of normal distribution can cause underestimation of the loss probability. Therefore, once the statistical significance of the q-Gaussian distributions is verified, this can be used for risk estimations more efficiently than the normal distribution.

In order to check whether the parameter estimates above are statistically significant, we calculate p-values of Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests (see Sec. 3.1.12). For the parameter estimates $(\hat{q}, \hat{\mu}_q, \hat{\sigma}_q)$,

[2] We select 30 currency pairs: AUD/JPY, BRL/JPY, CAD/JPY, CHF/JPY, EUR/AUD, EUR/BRL, EUR/CAD, EUR/CHF, EUR/GBP, EUR/JPY, EUR/MXN, EUR/NZD, EUR/SGD, EUR/USD, EUR/ZAR, GBP/JPY, MXN/JPY, NZD/JPY, SGD/JPY, USD/AUD, USD/BRL, USD/CAD, USD/CHF, USD/GBP, USD/JPY, USD/MXN, USD/NZD, USD/SGD, USD/ZAR and ZAR/JPY.

Table 5.3 ISO 4217 code, country and currency

Code	Country	Currency
AUD	Australia	Australian dollar
BRL	Brazil	Brazilian real
CAD	Canada	Canadian dollar
CHF	Switzerland	Swiss franc
GBP	United Kingdom	British pound
JPY	Japan	Japanese yen
MXN	Mexico	Mexican peso
NZD	New Zealand	New Zealand dollar
SGD	Singapore	Singapore dollar
USD	United States of America	United States dollar
ZAR	South Africa	South African Rand

the complementary cumulative distribution of the q-Gaussian is given by

$$F_1(x;\hat{q},\hat{\mu},\hat{\sigma}_q) = \begin{cases} \frac{1}{2}\left(1-\beta\left(\frac{\frac{\hat{q}-1}{(3-\hat{q})\hat{\sigma}_q^2}(x-\hat{\mu}_q)^2}{1-\frac{\hat{q}-1}{(3-\hat{q})\hat{\sigma}_q^2}(x-\hat{\mu}_q)^2};\frac{1}{2},\frac{1}{\hat{q}-1}-\frac{1}{2}\right)\right) & (x \geq \hat{\mu}_q) \\ \frac{1}{2}\left(1+\beta\left(\frac{\frac{\hat{q}-1}{(3-\hat{q})\hat{\sigma}_q^2}(x-\hat{\mu}_q)^2}{1-\frac{\hat{q}-1}{(3-\hat{q})\hat{\sigma}_q^2}(x-\hat{\mu}_q)^2};\frac{1}{2},\frac{1}{\hat{q}-1}-\frac{1}{2}\right)\right) & (x < \hat{\mu}_q) \end{cases} \tag{5.31}$$

When we let $F_T(x)$ be an empirical complementary cumulative distribution computed from T observations $\{r(0),\ldots,r(T-1)\}$ of log-returns, the distance between the empirical and assumed distributions is

$$z = \sup_{0\leq s\leq T-1} \sqrt{T}|F_T(r(s)) - F_1(r(s);\hat{q},\hat{\mu}_q,\hat{\sigma}_q)|\sqrt{\psi(F(r(s)))},$$

where T is the data length, $r(s)$ is a log-return of the currency and $\psi(u)$ is a weight function ($\psi(u)=1$ in the KS test or $\psi(u)=\frac{1}{u(1-u)}$ in the AD test). Then we obtain p-value. Table 5.5 shows p-values of q-Gaussian distributions fitted to the data of daily log-returns. Concerning the KS test, all the p-values are larger than 0.1. This means that the hypotheses that the log-returns obey the q-Gaussian are not rejected for all pairs with 10 % significance level if we do not focus on the tails. Therefore, it can be said that the log-returns of the exchange rates obey q-Gaussian distributions as a whole.

In the case of the AD test, however, the p-values of 9 pairs (AUD/JPY, EUR/AUD, EUR/BRL, EUR/MXN, EUR/NZD, NZD/JPY, USD/AUD, USD/CHF and USD/MXN) are less than 0.1. This means that the difference between the empirical and assumed distribution at tails is too large to say that the data are drawn from assumed distribution. Therefore, parameters estimated for the 9 pairs are not statistically significant, and cannot be used for risk assessment. One of the rea-

Table 5.4 Parameter estimates of the q-Gaussian for 30 currency pairs

Pair	$\hat{q}$	$\hat{\mu}_q$	$\hat{\sigma}_q$
AUD/JPY	1.609419	0.000704	0.006200
BRL/JPY	1.473631	0.000636	0.008481
CAD/JPY	1.445925	0.000306	0.006873
CHF/JPY	1.435333	0.000188	0.005205
EUR/AUD	1.435251	−0.000381	0.004816
EUR/BRL	1.460339	−0.000501	0.006275
EUR/CAD	1.258671	−0.000196	0.005356
EUR/CHF	1.705258	−0.000081	0.001919
EUR/GBP	1.392350	−0.000017	0.003845
EUR/JPY	1.456122	0.000187	0.005577
EUR/MXN	1.391250	−0.000068	0.005402
EUR/NZD	1.351964	−0.000393	0.005844
EUR/SGD	1.305555	−0.000130	0.004026
EUR/USD	1.343271	0.000027	0.005140
EUR/ZAR	1.328723	−0.000054	0.007293
GBP/JPY	1.499094	0.000041	0.005660
MXN/JPY	1.464356	0.000213	0.006993
NZD/JPY	1.532771	0.000522	0.007324
SGD/JPY	1.449534	0.000237	0.004899
USD/AUD	1.452021	−0.000513	0.006218
USD/BRL	1.473395	−0.000533	0.006568
USD/CAD	1.390452	−0.000229	0.004998
USD/CHF	1.326300	−0.000077	0.005608
USD/GBP	1.341826	−0.000043	0.005027
USD/JPY	1.359103	−0.000068	0.005070
USD/MXN	1.538956	−0.000307	0.004136
USD/NZD	1.373616	−0.000351	0.007173
USD/SGD	1.404451	−0.000223	0.002616
USD/ZAR	1.326522	−0.000068	0.008779
ZAR/JPY	1.400170	0.000125	0.009754

sons seems to be that asymmetry is seen in the empirical distribution for the log-returns of the exchange rates for the power or the economic circumstances between nations although the q-Gaussian is symmetry. This could result in discrepancy at tails between the empirical and the assumed distribution, which means that we need to assume another distribution which can express asymmetry. In Sect. 5.7, we treat another distribution, the Pearson type IV, which does not only include q-Gaussian distribution for $1 < q < 3$, but also has a parameter to express skewness.

Table 5.5 The p-values of q-Gaussian distributions fitted to the data of daily log-returns

Pair	KS	AD	Pair	KS	AD
AUD/JPY	0.202516	0.014810	GBP/JPY	0.786078	0.894253
BRL/JPY	0.259701	0.130124	MXN/JPY	0.590475	0.250290
CAD/JPY	0.782789	0.221188	NZD/JPY	0.109631	0.033321
CHF/JPY	0.912204	0.902014	SGD/JPY	0.749587	0.380528
EUR/AUD	0.821304	0.000398	USD/AUD	0.703012	0.004203
EUR/BRL	0.530867	0.081283	USD/BRL	0.542118	0.172771
EUR/CAD	0.750889	0.564035	USD/CAD	0.793130	0.779604
EUR/CHF	0.196867	0.421569	USD/CHF	0.918346	0.000000
EUR/GBP	0.356244	0.721617	USD/GBP	0.979804	0.441559
EUR/JPY	0.640597	0.213853	USD/JPY	0.192410	0.662888
EUR/MXN	0.473614	0.043808	USD/MXN	0.119169	0.004294
EUR/NZD	0.298500	0.033507	USD/NZD	0.735193	0.242241
EUR/SGD	0.912438	0.605466	USD/SGD	0.864110	0.759757
EUR/USD	0.608185	0.927556	USD/ZAR	0.824589	0.501781
EUR/ZAR	0.836994	0.513467	ZAR/JPY	0.625725	0.646715

5.7 Pearson Type IV Distribution

We applied q-Gaussian distribution in Sect. 5.6, and see that parameter estimates of the q-Gaussian are statistically significant as a whole, but not when we focus on tails. One reason seems to be that asymmetry of the empirical distributions. In this section, we introduce another distribution, the Pearson type IV distribution, which has fat-tails with skewness. The Pearson type IV distribution was derived as one of stationary distributions of a Pearson system [17]. Recently, there are several papers on an application of the Pearson type IV distribution to risk assessment [15, 20]. We estimate parameters for log-returns time series. Then, we discuss the validity of the assumption of the Pearson type IV distribution for the daily log-return of the exchange rates by using both KS and AD tests.

We consider the following distribution that we modify the q-Gaussian with a parameter α;

$$p(x) = A'_q \sqrt{B_q}\, \exp_q\left(-B_q(x-\mu_q)^2\right) \exp\left(\alpha \arctan\left(\sqrt{(q-1)B_q}(x-\mu_q)\right)\right), \tag{5.32}$$

$$A'_q = \frac{\sqrt{q-1}}{G\left(\frac{2}{q-1}-2, \alpha\right)}, \quad B_q = \frac{1}{(3-q)\sigma_q^2},$$

where $G(u, v)$ is Pearson function defined as

$$G(s,\lambda) = e^{-\frac{\pi\lambda}{2}} \int_0^\pi \sin^s\theta e^{\lambda\theta}\mathrm{d}\theta.$$

Equation (5.32) is called a Pearson type IV distribution, which we parametrise as a modification of the q-Gaussian. Here, we introduce the distribution as a stationary distribution of a certain diffusion process.

In Sect. 5.3, we have stated that the diffusion process Eq. (5.22),

$$\mathrm{d}u = -\nu u\mathrm{d}t + \sqrt{\gamma u^2 + \delta}\mathrm{d}W \ (\nu, \gamma, \delta > 0)\,,$$

results in the q-Gaussian as its stationary distribution. In order to consider a distribution that is better fitted to log-returns of the exchange rate than the q-Gaussian, we extend Eq. (5.22) to

$$\mathrm{d}u = (-\nu u + a)\mathrm{d}t + \sqrt{\gamma u^2 + \delta}\mathrm{d}W \ (\nu, \gamma, \delta > 0)\,. \tag{5.33}$$

Here, we show that the stationary solution of Eq. (5.33) can be described as Eq. (5.32). The corresponding Fokker–Planck equation of Eq. (5.33) is described as

$$\frac{\partial p(u,t)}{\partial t} = -\frac{\partial}{\partial u}\Big[(-\nu u + a)p(u,t)\Big] + \frac{1}{2}\frac{\partial^2}{\partial u^2}\Big[(\gamma u^2 + \delta)p(u,t)\Big]. \tag{5.34}$$

Similarly to Eq. (5.25) in Sect. 5.3, the stationary distribution $p(u)$ is the solution of the following equation:

$$-(-\nu u + a)p(u) + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}u}\Big[(\gamma u^2 + \delta)p(u)\Big] = 0,$$

which is transformed as

$$\frac{\mathrm{d}p(u)}{p(u)} = -\frac{2(\nu+\gamma)u}{\gamma u^2 + \delta}\mathrm{d}u + \frac{2a}{\gamma u^2 + \delta}\mathrm{d}u.$$

Then, this is integrated as

$$\ln p(u) = -\frac{\nu+\gamma}{\gamma}\ln(\gamma u^2 + \delta) + \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right) + C,$$

where C is a constant of integration. Therefore, the solution is given as

$$p(x) = C'\left(1 + \frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}} \exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)\right), \tag{5.35}$$

where C' is a normalisation constant. The constant C' is derived in Appendix C. Therefore, we obtain the stationary distribution of Eq. (5.33) as

$$p(u) = \frac{1}{G\left(\frac{2\nu}{\gamma}, \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\right)}\sqrt{\frac{\gamma}{\delta}}\left(1+\frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}} \exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)\right). \tag{5.36}$$

In order to treat the distribution as a modification of the q-Gaussian distribution, we make the transformation as $x - \mu_q = u$, $q = \frac{\nu+2\gamma}{\nu+\gamma}$, $\sigma_q^2 = \frac{\delta}{2\nu+\gamma}$, and $\alpha = \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}$. Then, we finally obtain Eq. (5.32).

The complementary cumulative distribution of Eq. (5.32) is described as

$$\begin{aligned} F_2(x) &= \Pr[X \geq x] \\ &= \frac{e^{-\frac{\pi\alpha}{2}}}{G\left(\frac{2}{q-1}-2, \alpha\right)} \int_0^{\frac{\pi}{2}-\arctan\left(\sqrt{(q-1)B_q(x-\mu_q)}\right)} (\sin\theta)^{\left(\frac{2}{q-1}-2\right)} e^{\alpha\theta}\mathrm{d}\theta. \end{aligned} \tag{5.37}$$

The detail derivation of the cumulative distribution is also shown in Appendix D.

5.7.1 Data Analysis

Using Eqs. (5.32) and (5.37), we perform parameter estimation procedures and statistical tests for the 30 currency pairs in the same manner as the q-Gaussian.

Table 5.6 shows parameter estimates obtained from empirical data. In addition to the 3 parameters q, μ_q and σ_q, α is estimated as a skewness of the distributions. The values of q and σ_q are almost equal to those in q-Gaussian distributions with differences by 0.1 % at most. However, μ_q is different from that of q-Gaussian distributions in each currency pair because the parameter α is added in the case of the Pearson type IV distribution. We note that, in Pearson type IV distribution, μ_q is not equivalent to average, and that it should be considered just as a location parameter. Therefore, we should focus on both μ_q and α when we want to know which currency pair is strong or weak.

We calculated the statistical significance of the parameter estimates. Table 5.7 shows p-values to indicate whether Pearson type IV distributions fitted to the empirical distributions for daily log-returns of the exchange rates. Concerning the KS test, all the p-values are larger than 0.2, which are better than those of q-Gaussian.

In the case of the AD test for the q-Gaussian distribution, p-values for only 9 pairs (AUD/JPY, EUR/AUD, EUR/BRL, EUR/MXN, EUR/NZD, NZD/JPY, USD/AUD, USD/CHF and USD/MXN) are less than 0.1, although, in the case of the AD

Table 5.6 Parameter estimates of the Pearson type IV distribution for 30 currency pairs

Pair	$\hat{q}$	$\hat{\mu}_q$	$\hat{\sigma}_q$	$\hat{\alpha}$
AUD/JPY	1.610139	0.001887	0.006151	−0.262934
BRL/JPY	1.472487	0.002940	0.008405	−0.437468
CAD/JPY	1.444235	0.001625	0.006851	−0.321471
CHF/JPY	1.434890	0.000991	0.005191	−0.261791
EUR/AUD	1.431137	−0.001284	0.004813	0.316844
EUR/BRL	1.455900	−0.002179	0.006243	0.438050
EUR/CAD	1.258280	−0.001598	0.005331	0.622461
EUR/CHF	1.704824	−0.000005	0.001920	−0.050019
EUR/GBP	1.393331	−0.000639	0.003830	0.293999
EUR/JPY	1.456696	0.001518	0.005531	−0.394483
EUR/MXN	1.389923	−0.001127	0.005385	0.355433
EUR/NZD	1.342089	−0.002884	0.005786	0.854961
EUR/SGD	1.306032	−0.000009	0.004025	−0.064110
EUR/USD	1.343607	0.000123	0.005139	−0.036981
EUR/ZAR	1.325985	−0.001973	0.007262	0.539210
GBP/JPY	1.500839	0.000770	0.005636	−0.199748
MXN/JPY	1.464645	0.001865	0.006939	−0.384782
NZD/JPY	1.532918	0.001876	0.007281	−0.278273
SGD/JPY	1.448921	0.001182	0.004878	−0.321406
USD/AUD	1.451001	−0.002055	0.006176	0.409805
USD/BRL	1.469904	−0.002161	0.006536	0.398716
USD/CAD	1.388414	−0.000927	0.004996	0.254392
USD/CHF	1.326647	0.000445	0.005602	−0.189920
USD/GBP	1.339521	−0.001051	0.005017	0.399280
USD/JPY	1.359006	0.000445	0.005065	−0.194651
USD/MXN	1.542045	−0.001543	0.004063	0.441634
USD/NZD	1.370606	−0.002314	0.007135	0.515685
USD/SGD	1.404469	−0.000672	0.002607	0.305095
USD/ZAR	1.325492	−0.002284	0.008735	0.517640
ZAR/JPY	1.401579	0.002084	0.009698	0.358533

test for Pearson type IV distributions, p-values of 5 pairs (EUR/AUD, EUR/BRL, EUR/CAD, USD/CHF and USD/GBP) are less than 0.1. Table 5.8 shows the average of p-value over 30 currency pairs with the two distributions in the KS and AD tests. This indicates that Pearson type IV distributions are better fitted to the empirical distributions calculated from the data than q-Gaussian distributions in average both as a whole or at the tails.

We also compare the results for the q-Gaussian and the Pearson type IV by Akaike Information Criterion (AIC). AIC is defined as

$$\mathrm{AIC} = -2L(\hat{\boldsymbol{\theta}}) + 2(K+1), \tag{5.38}$$

Table 5.7 The p-values of Pearson type IV distributions fitted to the data in the market

Pair	KS	AD	Pair	KS	AD
AUD/JPY	0.647416	0.733195	GBP/JPY	0.931475	0.886082
BRL/JPY	0.946681	0.402711	MXN/JPY	0.993608	0.994554
CAD/JPY	0.882842	0.995040	NZD/JPY	0.290311	0.721085
CHF/JPY	0.998742	0.771110	SGD/JPY	0.876989	0.935569
EUR/AUD	0.861370	0.016699	USD/AUD	0.759749	0.204882
EUR/BRL	0.892361	0.056225	USD/BRL	0.959783	0.348942
EUR/CAD	0.959756	0.035925	USD/CAD	0.651646	0.614915
EUR/CHF	0.261394	0.325228	USD/CHF	0.965324	0.000000
EUR/GBP	0.446183	0.951811	USD/GBP	0.999893	0.090961
EUR/JPY	0.978599	0.989188	USD/JPY	0.218458	0.711086
EUR/MXN	0.563741	0.314231	USD/MXN	0.983102	0.454872
EUR/NZD	0.778930	0.987520	USD/NZD	0.960887	0.828784
EUR/SGD	0.932922	0.561600	USD/SGD	0.956389	0.246523
EUR/USD	0.622775	0.927162	USD/ZAR	0.850672	0.884366
EUR/ZAR	0.991990	0.970829	ZAR/JPY	0.806047	0.988715

Table 5.8 The p-value averages with the two distributions in the KS and AD tests for the 30 currency pairs

Distribution	KS	AD
q-Gaussian	0.603903	0.372297
Pearson type IV	0.799001	0.598327

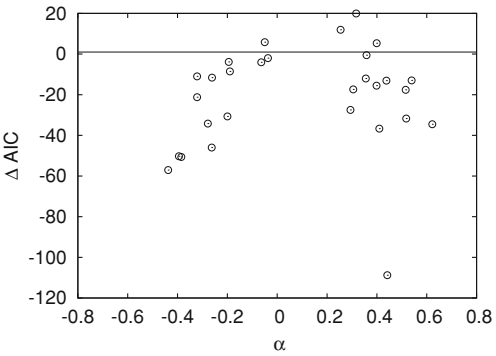


Fig. 5.3 Relation between the absolute value of the parameter estimates $\hat{\alpha}$ of the Pearson type IV and $\Delta\text{AIC} = \text{AIC}(\text{Pearson type IV}) - \text{AIC}(q\text{-Gaussian})$ for the 30 currency pairs. Each *unfilled circle* represents a relation for a currency pair

where $L(\theta)$ is the log-likelihood function in terms of the parameters $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ are the parameter estimates and K denotes the number of parameters. The smaller values of AIC mean that the model is better fitted to the given data.

Figure 5.3 shows the relation between the absolute values of the parameter estimate $\hat{\alpha}$ for the 30 currency pairs, and ΔAIC, the difference of AIC between the two distributions which we define as $\Delta\text{AIC} = \text{AIC}(\text{Pearson type IV}) - \text{AIC}(q-\text{Gaussian})$. There is negative correlation between $|\hat{\alpha}|$ and ΔAIC, and ΔAIC are larger than 0 in

the case that $|\hat{\alpha}|$ is small. This means that if an unconditional PDF for log-returns is skewed, the goodness-of-fit of the Pearson type IV distribution is better than one of the q-Gaussian distribution. Therefore, this indicates that a skewed parameter α plays an important role to explain the daily log-returns of the exchange rates.

5.7.2 Value at Risk and Expected Shortfall

We compare value at risk (VaR) and expected shortfall (ES) calculated from the q-Gaussian and the Pearson type IV distribution. VaR is one of the risk measures related to the loss probability. This is defined with confidence level c as

$$\Pr[x < -\mathrm{VaR}] = \int_{-\infty}^{-\mathrm{VaR}} p(x; \hat{\boldsymbol{\theta}})\mathrm{d}x = 1 - c, \tag{5.39}$$

where x is log-returns and $p(x; \hat{\boldsymbol{\theta}})$ is the assumed distribution with estimated parameters. Using the complementary cumulative distribution $F(x; \hat{\boldsymbol{\theta}})$, we have $1 - F(-\mathrm{VaR}; \hat{\boldsymbol{\theta}}) = 1 - c$. Then, VaR is calculated as an absolute value of the solution of $c = F(x; \hat{\boldsymbol{\theta}})$. c is generally chosen from the range of 95–99.5 %. This means that the losses larger than the amount of the VaR occur with probability $100(1 - c)$ % during the period for estimation. Expected shortfall (ES) is an alternative to value at risk that is more sensitive to the shape of the loss distribution in the tail of the distribution. The expected shortfall is defined as

$$ES_c = \frac{1}{c - 1} \int_c^1 q_F(\gamma)\mathrm{d}\gamma, \tag{5.40}$$

where $q_F(\gamma)$ is a quantile function of $F(x)$, which is defined as the inverse function of $F(x)$ so that $q_F(\gamma) = F^{-1}(\gamma)$. The expected shortfall is also expressed as

$$ES_c = VaR_c + \frac{1}{1 - c} \int_{-\infty}^{-VaR_c} (1 - F(x))\mathrm{d}x. \tag{5.41}$$

Table 5.9 shows 1 % ($c = 99$ %) VaR and ES of the 30 pairs calculated from q-Gaussian and Pearson type IV distributions with parameter estimates. We indicate $*$ in VaR and ES of pairs in which the p-values of the parameter estimates in the AD test are less than 1 %. Table 5.9 shows that VaRs and ESs in the two distributions differ by approximately 10 %, and especially the VaR and ES of the Pearson type IV seems to be larger than that of the q-Gaussian if α in the Pearson type IV is negative. For example, all pairs with JPY have negative α in the Pearson type IV, and the VaR

Table 5.9 1 % VaR and ES of the 30 pairs calculated from q-Gaussian and Pearson type IV distributions with estimated parameters

Pair	q-Gaussian		Pearson IV	
	VaR	ES	VaR	ES
AUD/JPY	0.035906	0.065811	0.041032	0.076342
BRL/JPY	0.035834	0.054405	0.040603	0.062731
CAD/JPY	0.027680	0.040877	0.030041	0.044836
CHF/JPY	0.020587	0.030209	0.021970	0.032532
EUR/AUD	0.019598*	0.028540*	0.018048	0.025815
EUR/BRL	0.026774	0.039757	0.023556	0.034043
EUR/CAD	0.016433	0.021230	0.015527	0.019848
EUR/CHF	0.015199	0.033776	0.015736	0.035032
EUR/GBP	0.014234	0.020215	0.013393	0.018842
EUR/JPY	0.022972	0.034329	0.025570	0.038826
EUR/MXN	0.020003	0.028153	0.018573	0.025776
EUR/NZD	0.020623	0.028038	0.017881	0.023533
EUR/SGD	0.013137	0.017549	0.013250	0.017726
EUR/USD	0.017536	0.023884	0.017643	0.024061
EUR/ZAR	0.024426	0.032708	0.022542	0.029706
GBP/JPY	0.025606	0.039958	0.027418	0.043281
MXN/JPY	0.029302	0.044001	0.032640	0.049815
NZD/JPY	0.035238	0.057390	0.039148	0.064664
SGD/JPY	0.019848	0.029573	0.021601	0.032556
USD/AUD	0.026131*	0.038480*	0.023317	0.033615
USD/BRL	0.028753	0.043245	0.025415	0.037261
USD/CAD	0.018656	0.026201	0.017691	0.024588
USD/CHF	0.018752*	0.025179*	0.019264*	0.025989*
USD/GBP	0.017178	0.023366	0.016111	0.021647
USD/JPY	0.017822	0.024514	0.018418	0.025467
USD/MXN	0.020790*	0.033973*	0.017369	0.027666
USD/NZD	0.026059	0.035908	0.023639	0.031911
USD/SGD	0.010103	0.014562	0.009453	0.013488
USD/ZAR	0.029302	0.039085	0.027180	0.035720
ZAR/JPY	0.036418	0.051292	0.029778	0.042931

* Represents that p-value of the KS or AD test is less than 1 %

and ES in the Pearson type IV are all larger than those in the q-Gaussian. Therefore, risk assessment with the q-Gaussian could result in underestimation of risks when the historical distribution is skewed sufficiently.

5.8 Conclusion

In this chapter, we applied the q-Gaussian and Pearson type IV distributions to assess currency risks of the foreign exchange market. In the hypothesis that log-returns of foreign exchange rates obey the q-Gaussian distribution, we performed

parameter estimation procedures for 30 currency pairs. In order to check whether the estimated parameters are statistically significant, we calculated p-values of two types of statistical test, Kolmogorov-Smirnov (KS) test [18, 19] and Anderson-Darling (AD) test [1].

The parameter q is estimated in the range of 1.3–1.7, which indicates that empirical distributions for the daily log-return of the exchange rates have fat-tails, and a risk assessment with normal distribution could result in underestimation of risk. We revealed that all p-values in the KS test are larger than 0.1, though p-values of 9 currency pairs were less than 0.1 in the AD test. This means that the log-returns obey the q-Gaussian as a whole, but do not always obey at tails.

We treated another distribution for risk assessment, the Pearson type IV distribution. We parametrise the distribution as a modification of the q-Gaussian with skewness and performed parameter estimations procedures for log-returns of the exchange rates. We revealed that the Pearson type IV distributions are better-fitted to the empirical distributions. All the p-values in the KS test were larger than 0.2, and p-values for only 5 pairs were less than 0.1 in the AD test. On average, p-values in both KS and AD tests with the Pearson type IV are larger than those with the q-Gaussian. This indicates that skew of the empirical distributions should be taken into consideration, and the Pearson type IV distribution is a more efficient model for risk assessment than the q-Gaussian.

We calculated 1 % value at risk (VaR) and expected shortfall (ES) of the currency pairs with the two distributions and compared the values. We found that VaR with the Pearson type IV are different from those with the q-Gaussian approximate of about 10 %. This indicates that even the assumption of the q-Gaussian distributions could result in underestimation of loss probability than the assumption of the Pearson type IV distribution.

Acknowledgments The author shows his sincere gratitude to Mr. Takashi Isogai (Bank of Japan) for his fruitful suggestions.

Appendix A: Derivation of q-Gaussian Distribution

Consider the maximisation of the Tsallis entropy

$$S_q[p] = \int_{-\infty}^{\infty} p(x) \ln_q \left(\frac{1}{p(x)} \right) \mathrm{d}x = \frac{1 - \int_{-\infty}^{\infty} p(x)^q \mathrm{d}x}{q - 1}, \tag{5.42}$$

under constraints

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1, \tag{5.43}$$

$$\int_{-\infty}^{\infty} x P_q(x)\mathrm{d}x = \mu_q, \tag{5.44}$$

$$\int_{-\infty}^{\infty} (x-\mu_q)^2 P_q(x)\mathrm{d}x = \sigma_q^2, \tag{5.45}$$

where μ_q is q-average and σ_q^2 q-variance, which are calculated with an escort probability $P_q(x) = p(x)^q / \int p(x)^q \mathrm{d}x$.

This optimisation problem can be solved by using the Lagrangian multipliers,

$$H = S_q[p] + \lambda_1 \left(\int_{-\infty}^{\infty} p(x)\mathrm{d}x - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x(p(x))^q \mathrm{d}x - \mu_q \int_{-\infty}^{\infty} (p(x))^q \mathrm{d}x \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu_q)^2 (p(x))^q \mathrm{d}x - \sigma_q^2 \int_{-\infty}^{\infty} (p(x))^q \mathrm{d}x \right), \tag{5.46}$$

where λ_i $(i = 1, 2, 3)$ are Lagrangian multipliers.

When $p(x)$ maximises H, we have

$$\frac{\partial H}{\partial p} = \int_{-\infty}^{\infty} \left(q(p(x))^{q-1} \left(\frac{1}{q-1} + \lambda_2 \left(x - \mu_q\right) - \lambda_3 \left(\left(x-\mu_q\right)^2 - \sigma_q^2 \right) \right) + \lambda_1 \right) \mathrm{d}x = 0.$$

This requires that the integrand is 0 for all x, so we have

$$q(p(x))^{q-1} \left(\frac{1}{q-1} + \lambda_2 \left(x - \mu_q\right) - \lambda_3 \left(\left(x-\mu_q\right)^2 - \sigma_q^2 \right) \right) + \lambda_1 = 0. \tag{5.47}$$

From Eq. (5.47), we get

$$\begin{aligned} p(x) &= \left(\frac{q-1}{q} \lambda_1 \right)^{\frac{1}{1-q}} \left(1 + (1-q) \left(\lambda_2 \left(x - \mu_q\right) - \lambda_3 \left(\left(x-\mu_q\right)^2 - \sigma_q^2 \right) \right) \right)^{\frac{1}{1-q}} \\ &= \left(\frac{q-1}{q} \lambda_1 \right)^{\frac{1}{1-q}} \left(1 - (1-q)\lambda_3 \left(\lambda_2^2 + \sigma_q^2 \right) + (1-q)\lambda_3 \left(x - \mu_q + \lambda_2 \right)^2 \right)^{\frac{1}{1-q}} \\ &= \left(\frac{(q-1)\lambda_1}{q(1-(1-q)\lambda_3 \left(\lambda_2^2 + \sigma_q^2 \right))} \right)^{\frac{1}{1-q}} \left(1 + \frac{\lambda_3}{1-(1-q)\left(\lambda_2^2 + \sigma_q^2 \right)} (1-q) \left(x - \mu_q + \lambda_2 \right)^2 \right)^{\frac{1}{1-q}} \\ &= \lambda_1' \left(1 + \lambda_3' (1-q) \left(x - \lambda_2' \right)^2 \right)^{\frac{1}{1-q}}, \end{aligned} \tag{5.48}$$

where we put λ_1', λ_2' and λ_3' as

$$\lambda_1' = \left(\frac{(q-1)\lambda_1}{q(1-(1-q)\lambda_3\left(\lambda_2^2+\sigma_q^2\right))} \right)^{\frac{1}{1-q}},$$
$$\lambda_2' = \mu_q - \lambda_2,$$
$$\lambda_3' = \frac{\lambda_3}{1-(1-q)\left(\lambda_2^2+\sigma_q^2\right)}.$$

In the case of $q > 3$, $\int_{-\infty}^{\infty} p(x)\mathrm{d}x$ diverges. Therefore, q must be less than 3.

In the case of $1 < q < 3$, in which $p(x)$ has fat-tails, we want to determine λ_1', λ_2' and λ_3' so that $p(x)$ satisfies the conditions Eqs. (5.43) to (5.45).

From Eq. (5.43), we have

$$\int_{-\infty}^{\infty} \lambda_1' \left(1+\lambda_3'(1-q)\left(x-\lambda_2'\right)^2\right)^{\frac{1}{1-q}} \mathrm{d}x = 1.$$

Considering the transformation $t = \frac{1}{1+\lambda_3'(1-q)\left(x-\lambda_2'\right)^2}$, we get

$$\begin{aligned}
&\int_{-\infty}^{\infty} \lambda_1' \left(1+\lambda_3'(1-q)\left(x-\lambda_2'\right)^2\right)^{\frac{1}{1-q}} \mathrm{d}x \\
&= \int_0^1 \frac{\lambda_1'}{\sqrt{(1-q)\lambda_3'}} t^{\frac{1}{q-1}-\frac{3}{2}}(1-t)^{-\frac{1}{2}} \mathrm{d}t \\
&= \frac{\lambda_1'}{\sqrt{(1-q)\lambda_3'}} B\left(\frac{1}{q-1}-\frac{1}{2},\frac{1}{2}\right) = 1.
\end{aligned} \tag{5.49}$$

From Eq. (5.44), we have

$$\int_{-\infty}^{\infty} (x-\mu_q)\lambda_1' \left(1+\lambda_3'(1-q)\left(x-\lambda_2'\right)^2\right)^{\frac{q}{1-q}} \mathrm{d}x = 0.$$

Putting $x' = x - \lambda_2'$ into Eq. (5.8), we have

$$\int_{-\infty}^{\infty} (x'+\lambda_2'-\mu_q)\lambda_1' \left(1+\lambda_3'(1-q)x'^2\right)^{\frac{q}{1-q}} \mathrm{d}x' = 0.$$

Because $\int_{-\infty}^{\infty} x' \left(1+\lambda_3'(1-q)x'^2\right)^{\frac{q}{1-q}} \mathrm{d}x' = 0$ for the integrand is an odd function, we get

$$\int_{-\infty}^{\infty} (\lambda_2' - \mu_q)\lambda_1' \left(1+\lambda_3'(1-q)x'^2\right)^{\frac{q}{1-q}} \mathrm{d}x' = 0.$$

This means

$$\lambda_2' = \mu_q. \tag{5.50}$$

From Eq. (5.45), we have

$$\int_{-\infty}^{\infty} \left((x-\mu_q)^2 - \sigma_q^2\right) \lambda_1' \left(1+\lambda_3'(1-q)\left(x-\lambda_2'\right)^2\right)^{\frac{q}{1-q}} \mathrm{d}x = 0.$$

By the same transformation of Eq. (5.45), we have

$$\left((1-q)\,\lambda_3'\right)^{-\frac{3}{2}} B\left(\frac{q}{q-1} - \frac{3}{2}, \frac{3}{2}\right) - \left((1-q)\,\lambda_3'\right)^{-\frac{1}{2}} \sigma_q^2 B\left(\frac{q}{q-1} - \frac{1}{2}, \frac{1}{2}\right) = 0.$$

Therefore, we obtain

$$\begin{aligned}
\lambda_3' &= \frac{1}{(1-q)\,\sigma_q^2} \frac{B\left(\frac{q}{q-1} - \frac{3}{2}, \frac{3}{2}\right)}{B\left(\frac{q}{q-1} - \frac{1}{2}, \frac{1}{2}\right)} \\
&= \frac{1}{(1-q)\,\sigma_q^2} \frac{\Gamma\left(\frac{q}{q-1} - \frac{3}{2}\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{q}{q-1}\right)} \frac{\Gamma\left(\frac{q}{q-1}\right)}{\Gamma\left(\frac{q}{q-1} - \frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&= -\frac{1}{\sigma_q^2(3-q)},
\end{aligned} \tag{5.51}$$

where $\Gamma(x)$ is the gamma function defined as

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \mathrm{d}t,$$

and we use the equality $a\Gamma(a) = \Gamma(a+1)$. From Eqs. (5.49) to (5.51), λ_1' is determined as

$$\lambda_1' = \frac{1}{\sigma_q^2 B\left(\frac{q}{q-1} - \frac{1}{2}, \frac{1}{2}\right)} \sqrt{\frac{q-1}{3-q}}. \tag{5.52}$$

Consequently, we have the q-Gaussian distribution for $1 < q < 3$,

$$p(x) = \frac{1}{B\left(\frac{q}{q-1} - \frac{1}{2}, \frac{1}{2}\right)} \sqrt{\frac{q-1}{(3-q)\sigma_q^2}} \left(1 + \frac{q-1}{(3-q)\sigma_q{}^2}(x-\mu_q)^2\right)^{\frac{1}{1-q}}. \tag{5.53}$$

Equation (5.53) can also be described as

$$p(x) = A_q \sqrt{B_q} \exp_q\left(-B_q(x-\mu_q)^2\right), \tag{5.54}$$

$$A_q = \frac{\sqrt{q-1}}{B\left(\frac{q}{q-1} - \frac{1}{2}, \frac{1}{2}\right)}, \qquad B_q = \frac{1}{(3-q)\sigma_q^2}.$$

Appendix B: Complementary Cumulative Distribution of the q-Gaussian

Let us derive the complementary cumulative distribution of the q-Gaussian,

$$F(x) = \int_x^\infty A_q \sqrt{B_q} (1 - (1-q) B_q (x' - \mu_q)^2)^{\frac{1}{1-q}} \mathrm{d}x'. \tag{5.55}$$

In the case of $x \geq \mu_q$, consider the following transformation:

$$t = \frac{1}{1 - (q-1) B_q (x - \mu_q)^2}.$$

Then, we have

$$F(x) = \frac{A_q}{2\sqrt{q-1}} \int_0^{\frac{1}{1-(1-q)B_q(x-\mu_q)^2}} t^{\frac{1}{q-1} - \frac{2}{3}} (1-t)^{-\frac{1}{2}} \mathrm{d}t. \tag{5.56}$$

Transforming $s = 1 - t$ with $\mathrm{d}s = -\frac{1}{2\sqrt{(q-1)B_q}} t^{-\frac{3}{2}} (1-t)^{-\frac{1}{2}} \mathrm{d}t$, we get

$$
\begin{aligned}
F(x) &= \frac{A_q}{2\sqrt{q-1}} \int_{\frac{(q-1)B_q(x-\mu_q)^2}{1-(1-q)B_q(x-\mu_q)^2}}^{1} s^{-\frac{1}{2}}(1-s)^{\frac{1}{q-1}-\frac{3}{2}}\mathrm{d}s \\
&= \frac{A_q}{2\sqrt{q-1}}\left(B\left(\frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right) - \int_{0}^{\frac{(q-1)B_q(x-\mu_q)^2}{1-(1-q)B_q(x-\mu_q)^2}} s^{\frac{1}{2}}(1-s)^{\frac{1}{q-1}-\frac{1}{2}}\mathrm{d}s \right) \\
&= \frac{1}{2}\left(1 - \frac{1}{B\left(\frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right)} \int_{0}^{\frac{(q-1)B_q(x-\mu_q)^2}{1-(1-q)B_q(x-\mu_q)^2}} s^{\frac{1}{2}}(1-s)^{\frac{1}{q-1}-\frac{1}{2}}\mathrm{d}s \right),
\end{aligned}
$$

where we use $A_q = \frac{\sqrt{q-1}}{B\left(\frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right)}$. Therefore, we have

$$
\begin{aligned}
F(x) &= \frac{1}{2}\left(1 - \beta\left(\frac{(q-1)B_q(x-\mu_q)^2}{1+(q-1)B_q(x-\mu_q)^2}; \frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right)\right) \\
&= \frac{1}{2}\left(1 - \beta\left(\frac{\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2}{1+\frac{q-1}{(3-q)\sigma_q^2}B_q(x-\mu_q)^2}; \frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right)\right). \qquad (5.57)
\end{aligned}
$$

In the case of $x \leq \mu_q$, using the same transformation $t = \frac{1}{1-(q-1)B_q(x-\mu_q)^2}$ ($\mathrm{d}x' = \frac{1}{2\sqrt{(q-1)B_q}} t^{-\frac{3}{2}}(1-t)^{-\frac{1}{2}}\mathrm{d}t$), we obtain

$$
\begin{aligned}
F(x) &= \int_{x}^{\infty} A_q\sqrt{B_q}(1-(1-q)B_q(x'-\mu_q)^2)^{\frac{1}{1-q}}\mathrm{d}x' \\
&= 1 - \int_{-\infty}^{x} A_q\sqrt{B_q}(1-(1-q)B_q(x'-\mu_q)^2)^{\frac{1}{1-q}}\mathrm{d}x' \\
&= 1 - \frac{A_q}{2\sqrt{q-1}} \int_{0}^{\frac{1}{1-(1-q)B_q(x-\mu_q)^2}} t^{\frac{1}{q-1}-\frac{2}{3}}(1-t)^{-\frac{1}{2}}\mathrm{d}t \\
&= 1 - \frac{1}{2}\left(1 - \beta\left(\frac{(q-1)B_q(x-\mu_q)^2}{1+(q-1)B_q(x-\mu_q)^2}; \frac{1}{2}, \frac{1}{q-1}-\frac{1}{2}\right)\right)
\end{aligned}
$$

$$= \frac{1}{2}\left(1 + \beta\left(\frac{(q-1)B_q(x-\mu_q)^2}{1+(q-1)B_q(x-\mu_q)^2}; \frac{1}{2}, \frac{1}{q-1} - \frac{1}{2}\right)\right)$$
$$= \frac{1}{2}\left(1 + \beta\left(\frac{\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2}{1+\frac{q-1}{(3-q)\sigma_q^2}(x-\mu_q)^2}; \frac{1}{2}, \frac{1}{q-1} - \frac{1}{2}\right)\right). \tag{5.58}$$

Appendix C: Derivation of the Normalisation Constant of Pearson Type IV Distribution

Let us determine the normalisation constant of Pearson type IV distribution,

$$p(u) = C'\left(1 + \frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}} \exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)\right). \tag{5.59}$$

Because this is normalised ($\int_{-\infty}^{\infty} p(u)\mathrm{d}u = 1$), we have

$$C'^{-1} = \int_{-\infty}^{\infty}\left(1 + \frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}} \exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)\right)\mathrm{d}u. \tag{5.60}$$

By using the transformation $\theta = \arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)$, Eq. (5.60) is transformed as

$$C'^{-1} = \sqrt{\frac{\delta}{\gamma}}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(\cos\theta)^{\frac{2\nu}{\gamma}}\exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\theta\right)\mathrm{d}\theta.$$

Furthermore, putting $\theta = -\theta'$ and $\theta' = \frac{\pi}{2} - \phi$, we have

$$C'^{-1} = \sqrt{\frac{\delta}{\gamma}}\int_{0}^{\pi}(\sin\phi)^{\frac{2\nu}{\gamma}}\exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\left(\phi - \frac{\pi}{2}\right)\right)\mathrm{d}\phi$$
$$= \sqrt{\frac{\delta}{\gamma}}G\left(\frac{2\nu}{\gamma}, \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\right). \tag{5.61}$$

Thus, we obtain the stationary distribution of Eq. (5.33) as

$$p(u) = \frac{1}{G\left(\frac{2\nu}{\gamma}, \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\right)}\sqrt{\frac{\gamma}{\delta}}\left(1+\frac{\gamma}{\delta}u^2\right)^{-\frac{\nu+\gamma}{\gamma}} \exp\left(\frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}\arctan\left(\sqrt{\frac{\gamma}{\delta}}u\right)\right). \tag{5.62}$$

Therefore, in order to treat the distribution as a modification of the q-Gaussian distribution, we make the transformation as $x-\mu_q = u$, $q = \frac{\nu+2\gamma}{\nu+\gamma}$, $\sigma_q^2 = \frac{\delta}{2\nu+\gamma}$ and $\alpha = \frac{2a}{\gamma}\sqrt{\frac{\gamma}{\delta}}$. Then, we finally obtain

$$p(x) = A'_q\sqrt{B_q}\exp_q\left(-B_q(x-\mu_q)^2\right)\exp\left(\alpha\arctan\left(\sqrt{(q-1)B_q}(x-\mu_q)\right)\right), \tag{5.63}$$

$$A'_q = \frac{\sqrt{q-1}}{G\left(\frac{2}{q-1}-2,\alpha\right)}, \quad B_q = \frac{1}{(3-q)\sigma_q^2}.$$

Appendix D: Derivation of the Cumulative Distribution Function of Pearson Type IV Distribution

Let us derive the complementary cumulative distribution of Eq. (5.63). Using the same transformation of $\theta = \arctan\left(\sqrt{(q-1)B_q}(x'-\mu_q)\right), \theta' = -\theta$ and $\theta' = \frac{\pi}{2}-\phi$, we obtain its complementary cumulative distribution as

$$\begin{aligned}
F_2(x) &= \int_x^\infty A'_q\sqrt{B_q}\exp_q\left(-B_q(x'-\mu_q)^2\right)\exp\left(\alpha\arctan\left(\sqrt{(q-1)B_q}(x'-\mu_q)\right)\right)\mathrm{d}x' \\
&= \frac{1}{G\left(\frac{2}{q-1}-2,\alpha\right)}\int_{\arctan\left(\sqrt{(q-1)B_q}(x-\mu_q)\right)}^{\frac{\pi}{2}}(\cos\theta)^{\left(\frac{2}{q-1}-2\right)}e^{\alpha\theta}\mathrm{d}\theta \\
&\quad \left(\text{where we use } \theta = \arctan\left(\sqrt{(q-1)B_q}(x'-\mu_q)\right)\right) \\
&= \frac{1}{G\left(\frac{2}{q-1}-2,\alpha\right)}\int_0^{\frac{\pi}{2}-\arctan\left(\sqrt{(q-1)B_q}(x-\mu_q)\right)}(\sin\phi)^{\left(\frac{2}{q-1}-2\right)}e^{\alpha\left(\phi-\frac{\pi}{2}\right)}\mathrm{d}\phi \\
&\quad \left(\text{where we use } \theta' = -\theta \text{ and } \phi = \frac{\pi}{2}-\theta'\right) \\
&= \frac{e^{-\frac{\pi\alpha}{2}}}{G\left(\frac{2}{q-1}-2,\alpha\right)}\int_0^{\frac{\pi}{2}-\arctan\left(\sqrt{(q-1)B_q}(x-\mu_q)\right)}(\sin\phi)^{\left(\frac{2}{q-1}-2\right)}e^{\alpha\phi}\mathrm{d}\phi.
\end{aligned} \tag{5.64}$$

References

1. Anderson, T.W., Darling, D.A.: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Stat. **23**, 193–212 (1952)
2. Beck, C.: Recent development in superstatistics. Braz. J. Phys. **39**(2A), 357–363 (2009)
3. BIS Triennial Central Bank Survey: Foreign exchange turnover in April 2013: preliminary global results. URL http://www.bis.org/publ/rpfx13fx.pdf. Accessed 31 Jan 2014
4. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. J Econometrics **31**, 307–327 (1986)
5. Clark, P.: A subordinated stochastic process model with finite variance for speculative prices. Econometrica **41**, 135–155 (1973)
6. Cont, R., Sornette, D.: Convergent multiplicative processes repelled from zero: power laws and truncated power laws. J. Phys. I Fr. **7**, 431–444 (1997)
7. Drożdż, S., Kwapień, J., Oświ ecimka, P., Rak, R.: The foreign exchange market: return distributions, multifractality, anomalous multifractality and the Epps effect. New J. Phys. **12**, 105003 (2010)
8. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica **50**, 987–1007 (1982)
9. Forman, J.L., Sørensen, M.: The pearson diffusions- a class of statistically tractable diffusion processes. Scand. J. Stat. **35**, 438–465 (2008)
10. Gabaix, X., Gopikrishnan, P., Plerou, V., Stanley, H.E.: Understanding the cubic and half-cubic laws of financial fluctuations. Phys. A **324**, 1–5 (2003)
11. Gabaix, X., Gopikrishnan, P., Plerou, V., Stanley, H.E.: A theory of power-law distributions in financial market fluctuations. Nature **423**, 267–270 (2003)
12. Gell-Mann, M., Tsallis, C.: Nonextensive Entropy, Interdisciplinary Applications. Oxford University Press, Oxford (2004)
13. de Haan, L., Reskick, S.I., Rootzén, H., de Varis, C.G.: Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. Stochast. Process. Appl. **32**, 213–224 (1989)
14. Mandelbrot, B., Taylor, H.M.: On the distribution of stock price differences. Oper. Res. **15**, 1057–1062 (1967)
15. Nagahara, Y.: The PDF and CF of Pearson type IV distribution and the ML estimation of he parameters. Stat. Probab. Lett. **43**(3), 251–264 (1999)
16. PACIFIC Exchange Rate Service: URL http://fx.sauder.ubc.ca/data.html. Accessed 10 Feb 2012
17. Pearson, K.: Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. Philos. Trans. R. Soc. Lond. **186**, 343–414 (1895)
18. Smirnov, N.: On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin methématiques de l'Université de Moscou 2, fasc. 2 (1939)
19. Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. **19**, 279–281 (1948)
20. Stavroyiannis, S., Makris, I., Nikolaidis, V., Zarangas, L.: Econometric modeling and value-at-risk using the Pearson type-IV distribution. Int. Rev. Financ. Anal. **22**, 10–17 (2012) http://dx.doi.org/10.1016/j.irfa.2012.02.003
21. Takayasu, H., Sato, A.-H., Takayasu, M.: Stable infinite variance fluctuations in randomly amplified Langevin systems. Phys. Rev. Lett. **79**, 966–969 (1997)
22. Thistleton, W.J., Marsh, J.A., Nelson, K., Tsallis, C.: Generalized Box-Müller method for generating q-Gaussian random deviates. IEEE Trans. Inf. Theor. **53**(12), 4805–4810 (2007)

Chapter 6
Segmentation Study of Foreign Exchange Market

Abstract This chapter explains a recursive segmentation procedure under normal distribution assumptions. The Akaike information criterion between independently identically distributed Gaussian samples and two successive segments drawn from different Gaussian distributions is used as a discriminator to segment time series. The Jackknife method is employed in order to evaluate a statistical significance level. This chapter shows univariate and multivariate cases. The proposed method is performed for artificial time series consisting of two segments with different statistics. Furthermore, log-return time series of currency exchange rates for 30 currency pairs for the period from January 4, 2001 to December 30, 2011 are divided into 11 segments with the proposed method. It is confirmed that some segment corresponds to historical events recorded as critical situations.

6.1 Introduction

Both a mixture of distributions and switching models provide good expressions for non-stationary time series. Specifically, it is powerful to employ statistical inference methods under the assumption of these models combining with a model selection such as a likelihood-ratio test [33]. This type of methods is called change-point detection.

Detecting and modelling structural change and break-point from time series are often needed when we consider an problem of socioeconomic-technological systems. There are successive studies on change-point detection including monographs [4, 6, 8, 13]. Giraitis and Leipus [16, 17] study a method to detect a change-point by means of power spectra. Hawkins [23, 24], Chen and Gupta [9], Mia and Zhao [32], Sen and Srivastava [42] among others, are also of interest.

There are two types of approaches to change-point detection. One is a regression analysis. Several segments with different coefficients between an explained variable and explanatory variables. This is called spanned regression or segmentation. The Chow test is a statistical test to determine whether the coefficients in two linear regressions on different data sets are equal [12]. Quandt and Ramsey [40] have

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_6, © Springer Japan 2014

developed estimation procedure for mixture of linear regression. Hansen's [22] test of model stability was based on a cumulative sum of the least squares residuals.

Another approach is to determine change-points from a time series based on some models. Markov [21] switching models, stochastic differential equations [43], Gaussian models [10] are often assumed for this purpose. The likelihood ratio or some test statistics are used to determine breakpoints from a uni-variate time series or multivariate time series.

Two types of approaches to divide time series into several segments. One is a local approach and another is a global approach. In the local approach, a binary segmentation procedure is recursively applied to segments. There are some termination conditions to decide whether the binary segmentation procedure is applied or not. It is also considered to use a penalty function to prevent too many segments to be generated. The global approach solves a nonlinear optimisation problems in terms of parameters in segments.

Kawahara and Sugiyama [26] propose a non-parametric method to detect change points from time series based on direct density-ratio estimation. More recently, a recursive entropic scheme to separate financial time series has been proposed [10]. Their method is parametric and uses the log-likelihood ratio test. Ducré-Robitaille et al. [14] compare several methods to detect change points. They segment artificial time series based on 8 methods; standard normal homogeneity test (SNHT) without trend [2], SNHT with trend [3], multiple linear regression (MLR) [45], two-phase regression (TPR) [15], Wilcoxon rank–sum (WRS) [25], sequential testing for equality of means (ST) [20], Bayesian approach without reference series [37, 38], and Bayesian approach with reference series [37, 38]. Karl and Williams [25] propose a method to find an adequate segment boundary based on Wilcoxon rank-sum test and investigate climatological time series data.

I further address some existing approaches to the problem of multiple change-point detection in multivariate time series. Ombao et al. [34] employed the SLEX (smooth localised complex exponentials) basis for time series segmentation, originally proposed by Ombao et al. [35]. The choice of SLEX basis leads to the segmentation of the time series, achieved via complexity-penalised optimisation. Lavielle and Teyssière [29] introduced a procedure based on penalised Gaussian log-likelihood as a cost function, where the estimator was computed via dynamic programming. Sato proposes Akaike Information Criterion of multivariate Gaussian models [41]. The performance of the method was tested on bi-variate examples. Vert and Bleakley [44] proposed a method for approximating multiple signals (with independent noise) via piecewise constant functions, where the change-point detection problem was reformulated as a penalised regression problem and solved by the group Lasso [47]. Note that Cho and Fryzlewicz [11] argued that Lasso-type penalties were sub-optimal for change-point detection.

It still remains a challenge for financial engineering. The segmentation method presented in this chapter is an attempt to provide some insights into the problem of finding transition points in financial time series, or more generally in multivariate financial data. We consider an application of a method to detect change-point for multivariate time series to the foreign exchange rates. Statistical properties of asset

price returns in stock markets have been extensively studied in the literature of finance and econophysics over the past two decades [5, 31]. By these studies, many stylised facts about the statistics of returns have been determined. An important property of prices statistics is that the probability distribution of stock returns exhibits a fat-tailed distribution [19, 30]. The presence of fat tails in the observed return distributions can be partially attributed to the non-stationarity of the underlying processes that are responsible for shaping the prices of assets in the market. Indeed, even if a time series representing asset returns is a mixture of two Gaussian processes that are flip-flopped in consecutive time intervals, the resulting distribution is not Gaussian.

Actually, if the volatility of the two Gaussian distributions differs significantly, one effectively detects a distribution that has a large kurtosis as if it had a fat-tail. More generally, almost any distribution, including fat-tailed distributions with power-like tails, can be obtained as a weighted composition of Gaussian random variables. Non-stationary time series can easily explain fat-tails in historical financial data. Therefore, a systematic treatment and control of non-stationary effects in a financial time series are important.

In this chapter, I propose a segmentation procedure for a multivariate time series under the assumption of local stationarity, which means that the corresponding time series are generated from different multivariate Gaussian distributions that are stationary in given time intervals. I adopt this scheme here to analyse multivariate data on foreign exchange rates. The proposed procedure is applied to the process of segmenting multiple daily log-return time series of currency exchange rates for selected 30 currency pairs for the period from January 4, 2001 to December 30, 2011.

This chapter is organised as follows. Sections 6.2 and 6.3 briefly explain the likelihood-ratio test for segmentation of both univariate and multivariate time series. Section 6.4 proposes the statistical test for segmentation of multivariate time series based on information criterion. Section 6.5 interprets the statistical error of test statistics from random matrix theory. Section 6.6 shows results of the proposed segmentation procedure with an artificial multivariate time series. Section 6.7 shows empirical analysis of the multivariate time series of daily foreign exchange rates. Section 6.8 is devoted to concluding remarks.

6.2 Likelihood-Ratio Test for Univariate Time Series

Suppose that there are T observations x_s ($s = 1, \ldots, T$). We assume that the time series consists of m locally stationary segments (statistics within each segment are assumed to be homogeneous) and that each segment is sampled from a Gaussian distribution with different mean and variance. How do we determine the unknown $m - 1$ segment boundaries t_i ($i = 1, \ldots, m - 1$).

Recently, Cheong et al. have considered a recursive segmentation scheme for one dimensional time series under a Gaussian assumption [10] in the context of financial time series analysis. Their procedure can be interpreted as a kind of hypothesis test between a null hypothesis and an alternative model.

Let $g(x; \mu, \sigma^2)$ be a Gaussian distribution:

$$g(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big], \tag{6.1}$$

Assuming that the observations x_s should be segmented at t, and that the observations on the left hand side are sampled from $g(x; \mu_L, \sigma_L^2)$, and that those on the right hand side are from $g(x; \mu_R, \sigma_R^2)$, we define likelihood functions:

$$L_1 = \prod_{s=1}^{T} g(x_s; \mu, \sigma^2), \tag{6.2}$$

$$L_2(t) = \prod_{s=1}^{t} g(x_s; \mu_L, \sigma_L^2) \prod_{s=t+1}^{T} g(x_s; \mu_R, \sigma_R^2). \tag{6.3}$$

The log-likelihood ratio test can be constructed from the logarithmic difference between L_1 and $L_2(t)$, defined as

$$\Delta(t) = \ln L_2(t) - \ln L_1. \tag{6.4}$$

Inserting Eqs. (6.2) and (6.3) into Eq. (6.4), we have

$$\Delta(t) = \sum_{s=1}^{t} \ln g(x_s; \mu_L, \sigma_L^2) + \sum_{s=t+1}^{T} \ln g(x_s; \mu_R, \sigma_R^2) - \sum_{s=1}^{T} \ln g(x_s; \mu, \sigma^2). \tag{6.5}$$

In general, if a random variable A is given and its distribution admits a PDF p, then from T random variables a_i $(i = 1, \ldots, T)$ the expected value of $f(A)$, where f is a function, (if exists) can be approximated as

$$\frac{1}{T}\sum_{i=1}^{T} f(a_i) \approx \mathrm{E}[f(A)] = \int_{-\infty}^{\infty} f(x')p(x')\mathrm{d}x'. \tag{6.6}$$

By using this approximation for T observations x_s, we obtain

$$\sum_{s=1}^{T} \ln g(x_s; \mu, \sigma^2) \approx T \int_{-\infty}^{\infty} g(x; \mu, \sigma^2) \ln g(x; \mu, \sigma^2)\mathrm{d}x = -\frac{T}{2}\ln(2\pi e\sigma^2). \tag{6.7}$$

Furthermore, the standard deviations σ, σ_L and σ_R can be estimated as the sample standard deviations defines as

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_{s=1}^{T}\Big(x_s - \frac{1}{T}\sum_{s'=1}^{T} x_{s'}\Big)^2}, \tag{6.8}$$

$$\hat{\sigma}_L = \sqrt{\frac{1}{t}\sum_{s=1}^{t}\Big(x_s - \frac{1}{t}\sum_{s'=1}^{t} x_{s'}\Big)^2}, \tag{6.9}$$

$$\hat{\sigma}_R = \sqrt{\frac{1}{T-t}\sum_{s=t+1}^{T}\Big(x_s - \frac{1}{T-t}\sum_{s'=t+1}^{T} x_{s'}\Big)^2}. \tag{6.10}$$

Therefore, $\Delta(t)$ is empirically calculated as,

$$\Delta(t) = T\ln\hat{\sigma} - t\ln\hat{\sigma}_L - (T-t)\ln\hat{\sigma}_R \geq 0. \tag{6.11}$$

$\Delta(t)$ can be used as an indicator to separate the observations into two parts. An adequate procedure to separate the observations is that we choose the boundary at t where $\Delta(t)$ is maximised,

$$t^* = \arg\max_t \Delta(t). \tag{6.12}$$

If $\max_t \Delta(t)$ is less than a threshold value Δ_c, then this procedure should be terminated.

This process is recursively applied to each segmented time series. After separate the time series into two parts, we also apply this procedure for each segment hierarchically. A multivariate version of this procedure will be seen in the next section.

6.3 Likelihood-Ratio Test for M-Dimensional Multiple Time Series

Let $\boldsymbol{x}(s) = (x_1(s), \ldots, x_M(s))^{\mathrm{T}}$ $(s = 1, \ldots, T)$ be the M-dimensional multiple time series. Let further us assume that the multivariate time series consists of m sequences sampled from m different multivariate Gaussian distributions. We further assume that a segment k follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}^{(k)} = (\mu_1^{(k)}, \ldots, \mu_M^{(k)})$ and a variance-covariance matrix $\boldsymbol{C}^{(k)}$,

$$p\big(\boldsymbol{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{C}^{(k)}\big) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{C}^{(k)}|^{1/2}} \exp\Big[-\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}[(\boldsymbol{C}^{(k)})^{-1}]_{ij}(x_i - \mu_i^{(k)})(x_j - \mu_j^{(k)})\Big]. \tag{6.13}$$

To determine the m stationary segments from the given T observations of multiple time series $\boldsymbol{x}(s)$, let us consider a recursive segmentation procedure based on the

likelihood-ratio test. The likelihood-ratio test is one of the most efficient tests of statistical hypotheses [33].

Let us consider a method to determine a segment boundary between two segments for a multivariate time series. The null hypothesis assumes that T observations are sampled from an *i.i.d* M-dimensional Gaussian distribution $p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C})$, and the alternative model assumes that left t successive observations are sampled from an *i.i.d* M-dimensional Gaussian distribution $p(\boldsymbol{x}; \boldsymbol{\mu}_L, \boldsymbol{C}_L)$ and that right $T-t$ successive observations are sampled from an *i.i.d* M-dimensional Gaussian distribution $p(\boldsymbol{x}; \boldsymbol{\mu}_R, \boldsymbol{C}_R)$. In this case, the test statistic $\Delta(t)$ consisting of likelihood values defined as

$$L_1 = \prod_{s=1}^{T} p(\boldsymbol{x}(s); \boldsymbol{\mu}, \boldsymbol{C}), \tag{6.14}$$

$$L_2(t) = \prod_{s=1}^{t} p(\boldsymbol{x}(s); \boldsymbol{\mu}_L, \boldsymbol{C}_L) \prod_{s=t+1}^{T} p(\boldsymbol{x}(s); \boldsymbol{\mu}_R, \boldsymbol{C}_R), \tag{6.15}$$

is given as

$$\Delta(t) = \ln L_2(t) - \ln L_1. \tag{6.16}$$

In the M-dimensional Gaussian case, the logarithmic likelihood-ratio $\Delta(t)$ is computed as

$$\Delta(t) = \frac{T}{2} \ln |\hat{\boldsymbol{C}}| - \frac{t}{2} \ln |\hat{\boldsymbol{C}}_L| - \frac{T-t}{2} \ln |\hat{\boldsymbol{C}}_R|, \tag{6.17}$$

since Eqs. (6.14) and (6.15) are approximated as

$$\ln L_1 \approx -\frac{T}{2} \ln |\hat{\boldsymbol{C}}| - \frac{TM}{2} \ln(2\pi) - \frac{TM}{2}, \tag{6.18}$$

$$\ln L_2(t) \approx -\frac{t}{2} \ln |\hat{\boldsymbol{C}}_L| - \frac{T-t}{2} \ln |\hat{\boldsymbol{C}}_R| - \frac{TM}{2} \ln(2\pi) - \frac{TM}{2}, \tag{6.19}$$

where $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{C}}_L$, and $\hat{\boldsymbol{C}}_R$ represent their maximum likelihood estimators (empirical variance-covariance matrix), respectively, defined as

$$\hat{C}_{ij} = \frac{1}{T} \sum_{s=1}^{T} \Big(x_i(s) - \hat{\mu}_i\Big)\Big(x_j(s) - \hat{\mu}_j\Big), \tag{6.20}$$

$$\hat{C}_{L,ij} = \frac{1}{t} \sum_{s=1}^{t} \Big(x_i(s) - \hat{\mu}_{L,i}\Big)\Big(x_j(s) - \hat{\mu}_{L,j}\Big), \tag{6.21}$$

$$\hat{C}_{R,ij} = \frac{1}{T-t}\sum_{s=t+1}^{T}\Big(x_i(s) - \hat{\mu}_{R,i}\Big)\Big(x_j(s) - \hat{\mu}_{R,j}\Big), \tag{6.22}$$

where

$$\hat{\mu}_i = \frac{1}{T}\sum_{s=1}^{T} x_i(s), \tag{6.23}$$

$$\hat{\mu}_{L,i} = \frac{1}{t}\sum_{s=1}^{t} x_i(s), \tag{6.24}$$

$$\hat{\mu}_{R,i} = \frac{1}{T-t}\sum_{s=t+1}^{T} x_i(s). \tag{6.25}$$

The derivation of Eqs. (6.18) and (6.19) is shown in Appendix A.

The spectrum of the log likelihood-ratio $\Delta(t)$ in terms of t has a maximum at some time denoted as t^*,

$$\Delta^* = \Delta(t^*) = \max_t \Delta(t). \tag{6.26}$$

The interpretation of this time point t^* is that it gives an optimal separation of the multiple time series into two statistically most distinct segments. This segmentation procedure can be used recursively to separate further the multiple time series into smaller segments. We do this iteratively until the iteration is terminated by a stopping condition. As the stopping condition, Cheong et al. assume a constant threshold value Δ_{th}. If $\Delta^* < \Delta_{th}$, then the recursive segmentation procedure is terminated and does not proceed any more. Wilks proposes the test statistics $2\Delta(t)$ should asymptotically follow χ-squared distribution with degree of freedom equal to the difference of the number of parameters between the alternative and null models: $(M^2+3M+2)/2$ [46].

6.4 Information Criterion Test for M-Dimensional Multiple Time Series

However, since the likelihood-ratio test is a kind of model selection problem, we need to use test statistics constructed from an information criterion. Assuming Akaike Information Criterion (AIC) [1] as the information criterion, I attempt to reconstruct the likelihood-ratio test. The AIC of a model with K model parameters $\boldsymbol{\theta}$ for T observations is defined as

$$AIC = -2L(\hat{\boldsymbol{\theta}}) + 2K, \tag{6.27}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood value of the model with the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. The AIC value AIC_1 for the null model expressed in Eq. (6.14) is given

by the *i.i.d* M-dimensional Gaussian distribution with T observations;

$$AIC_1 = T \ln |\hat{\boldsymbol{C}}| + TM \ln(2\pi) + TM + (M^2 + 3M + 2), \tag{6.28}$$

Similarly to AIC_1, the AIC value $AIC_2(t)$ for the alternative model is described as

$$\begin{aligned} AIC_2(t) = \; & t \ln |\hat{\boldsymbol{C}}_L| + (T - t) \ln |\hat{\boldsymbol{C}}_R| \\ & + TM \ln(2\pi) + TM + 2(M^2 + 3M + 2). \end{aligned} \tag{6.29}$$

Therefore, I propose that the test statistic $\Delta(t)$ based on information criterion may be modified as

$$\begin{aligned} \Delta_{AIC}(t) &= AIC_2(t) - AIC_1 \\ &= t \ln |\hat{\boldsymbol{C}}_L| + (T - t) \ln |\hat{\boldsymbol{C}}_R| - T \ln |\hat{\boldsymbol{C}}| + (M^2 + 3M + 2). \end{aligned} \tag{6.30}$$

The spectrum of the log likelihood-ratio $\Delta_{AIC}(t)$ in terms of t has a minimum at some time denoted as t^*,

$$\Delta_{AIC}(t^*) = \min_t \Delta_{AIC}(t). \tag{6.31}$$

If $\Delta_{AIC}(t^*) < 0$, then the multivariate time series is divided into two segments at t^*. Otherwise, the segmentation procedure is terminated. The interpretation of this time point t^* is that it gives an optimal separation of the multiple time series into two statistically most distinct segments.

As a termination condition, it is further necessary to introduce a statistical significance level. This idea behind this termination condition is as follows. The test statistic $\Delta(t)$ contains estimation error which is determined by the number of observations T, the segment boundary t^*, and sampled variance-covariances $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{C}}_L$, and $\hat{\boldsymbol{C}}_R$.

Assume that $\tilde{\boldsymbol{C}}$, $\tilde{\boldsymbol{C}}_L$, and $\tilde{\boldsymbol{C}}_R$ are denoted as Jackknife variance-covariance matrices computed from K Jackknife segments from multiple time series $\boldsymbol{x}(s)$ $(s = 1, \ldots, T)$. Let γ $(0 \leq \gamma \leq 1)$ be a ratio to determine the length of Jackknife segments. $\tilde{\boldsymbol{C}}_L$ are computed from Jackknife sequences $\boldsymbol{x}_L(s)$ $(s = \tau_k, \ldots, \tau_k + [t^*\gamma])$, where τ_k is randomly selected with the same probability for $1 \leq \tau_k \leq t - [t\gamma]$. $[\cdot]$ represents the largest integer which is less than or equal to $\cdot$. $\tilde{\boldsymbol{C}}_R$ is computed from Jackknife sequences $\boldsymbol{x}_R(s)$ $(s = \tau'_k, \ldots, \tau'_k + [(T - t^*)\gamma])$, where τ'_k is randomly selected with the same probability for $t^* + 1 \leq \tau'_k \leq T - [(T - t^*)\gamma]$. $\tilde{\boldsymbol{C}}$ is computed from both the Jackknife sequences $\boldsymbol{x}_L(s)$ $(s = \tau_k, \ldots, \tau_k + [t^*\gamma])$ and $\boldsymbol{x}_R(s)$ $(s = \tau'_k, \ldots, \tau'_k + [(T - t^*)\gamma])$. From these Jackknife values, we can compute the Jackknife test statistic:

$$\tilde{\Delta}_{AIC}(t^*) = t^* \ln |\tilde{\boldsymbol{C}}_L| + (T - t^*) \ln |\tilde{\boldsymbol{C}}_R| - T \ln |\tilde{\boldsymbol{C}}| + (M^2 + 3M + 2). \tag{6.32}$$

The estimation error of the test statistic $\Delta_{AIC}(t^*)$ is estimated from the Jackknife density of $\tilde{\Delta}_{AIC}(t^*)$. The p-value of the event where $\Delta_{AIC}(t^*) < 0$ is approximated as

$$\Pr\Big[\Delta_{AIC}(t^*) < 0\Big] \approx \frac{K\Big[\tilde{\Delta}_{AIC}(t^*) < 0\Big]}{K}, \tag{6.33}$$

where $K[\Delta_{AIC}(t^*) < 0]$ is the number of events where $\tilde{\Delta}_{AIC}(t^*) < 0$ is satisfied. We can use the probability $\Pr[\Delta_{AIC}(t^*) < 0]$ as the statistical significance level. Namely, the multivariate time series is divided into two segments at t^* if $\Pr[\Delta_{AIC}(t^*) < 0] > \alpha_{th}$, where α_{th} is the significance level.

6.5 Estimation Error

Let us now briefly discuss some issues related to the estimation error to the maximum likelihood estimators of variance-covariance matrix computed from T successive observations.

$$\begin{aligned}
\Delta_{AIC}(t) =& \; t\ln|\hat{\boldsymbol{C}}_L| + (T-t)\ln|\hat{\boldsymbol{C}}_R| - T\ln|\hat{\boldsymbol{C}}| + (M^2+3M+2) \\
=& \; t\sum_{i=1}^{M}\ln\lambda_i^{(L)} + (T-t)\sum_{i=1}^{M}\ln\lambda_i^{(R)} - T\sum_{i=1}^{M}\ln\lambda_i + (M^2+3M+2) \\
\approx& \; M\Big(t\int_0^\infty \rho_L(\lambda)\ln\lambda d\lambda + (T-t)\int_0^\infty \rho_R(\lambda)\ln\lambda d\lambda - T\int_0^\infty \rho(\lambda)\ln\lambda d\lambda\Big) \\
& + (M^2+3M+2),
\end{aligned} \tag{6.34}$$

where λ_i, $\lambda_i^{(L)}$, and $\lambda_i^{(R)}$ represent eigenvalues of the corresponding variance-covariance matrices: $\boldsymbol{C}$, $\boldsymbol{C}_L$, and $\boldsymbol{C}_R$, respectively. These matrices are estimated from data by using Eqs. (6.20), (6.21) and (6.22). The relationship between true variance-covariance matrix and sampled variance-covariance matrix describing the underlying correlations can be found using random matrix theory [7, 28, 36, 39].

In general, the eigenvalue distribution of a sample variance-covariance matrix depends on the ratio between the length of the data set T and the number of degree of freedom M. In particular, if $\boldsymbol{x}(s)$ $(s = 1, \ldots, T)$ describes M-multivariate Gaussian uncorrelated identically distributed random variables, the density of the eigenvalues of the sample variance-covariance is approximated by the Marčnko-Pastur density:

$$\rho(\lambda) = \begin{cases} \frac{T}{M}\frac{\sqrt{(\lambda-\lambda_-)(\lambda-\lambda_+)}}{2\pi\sigma^2\lambda} & (\lambda_- \le \lambda \le \lambda_+) \\ 0 & (\text{otherwise}) \end{cases}, \tag{6.35}$$

where $\lambda_\pm = \sigma^2\Big(1\pm\sqrt{\frac{M}{T}}\Big)$ and σ^2 is a scale factor related to the variance of individual degrees of freedom. At $M = T$, the last equation reduces to

$$\rho(\lambda) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{\frac{\lambda_+ - \lambda}{\lambda}} & (0 < \lambda \leq \lambda_+) \\ 0 & (\lambda > \lambda_+) \end{cases}, \tag{6.36}$$

where $\lambda_+ = 4\sigma^2$. In this case, the integrand in the last formula for $\Delta_{AIC}(t)$ becomes singular at $\lambda = 0$ and, in effect, $\Delta(t)$ is ill-defined. From the statistical point of view this means that $\Delta(t)$ is estimated with a huge statistical uncertainty when M approaches T and, thus, this makes it practically impossible to estimate t^* properly. This situation is even worse for $M/T > 1$, since then the density has a peak at $\lambda = 0$. The integral is well defined only if $T > M$. The same holds for each of the subsystems: $t^* > M$ and $T - t^* > M$ and, in effect, t^* is restricted to a certain range, from t_{min} to t_{max}, for which the two inequalities are fulfilled. So far, I have discussed the simplest case of uncorrelated *i.i.d* numbers. For a correlated multivariate time series, the situation is more complicated but, in general, one expects a finite window for t^*. Actually, for a mixture of two Gaussian distributions there are further limitations on t_{min} and t_{max}, which are related to the statistical significance of the separation of two different eigenvalues that can be made for a given sample variance-covariance matrix. Typically, to distinguish two eigenvalues of the variance-covariance matrix one needs $M > 3T$ [7].

6.6 Numerical Study

As test data, we generate artificial multiple time series consisting of four segments ($m = 4$). Each segment is an M-dimensional multivariate time series drawn from a normal distribution with a specific variance-covariance matrix. The test time series is generated from the following procedure:

$$\boldsymbol{x}(t) = \begin{cases} \boldsymbol{A}_1\boldsymbol{\xi}(t) & (1 \leq t \leq 100) \\ \boldsymbol{A}_2\boldsymbol{\xi}(t) & (101 \leq t \leq 200) \\ \boldsymbol{A}_3\boldsymbol{\xi}(t) & (201 \leq t \leq 300) \\ \boldsymbol{A}_4\boldsymbol{\xi}(t) & (301 \leq t \leq 400) \end{cases}, \tag{6.37}$$

where $\boldsymbol{x}(t)$ represents an M-dimensional column vector expressing dynamical variables at time t and $\boldsymbol{\xi}(t)$ is an M-dimensional column vector expressing random fluctuations drawn from *i.i.d* standard normal distributions. $\boldsymbol{A}$ is an $M \times M$ random matrix where each element sampled from a standard normal distribution. This procedure provides M-dimensional multiple time series with different variance-covariance matrices depending on segments. The variance-covariance matrix at each segments is given as

$$\boldsymbol{C}^{(k)} = \sum_{t=t_{k-1}}^{t_k-1} \boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t) = \boldsymbol{A}_k \sum_{t=t_{k-1}}^{t_k-1} \boldsymbol{\xi}(t)\boldsymbol{\xi}^{\mathrm{T}}(t)\boldsymbol{A}_k^{\mathrm{T}} = \boldsymbol{A}_k\boldsymbol{A}_k^{\mathrm{T}}, \tag{6.38}$$

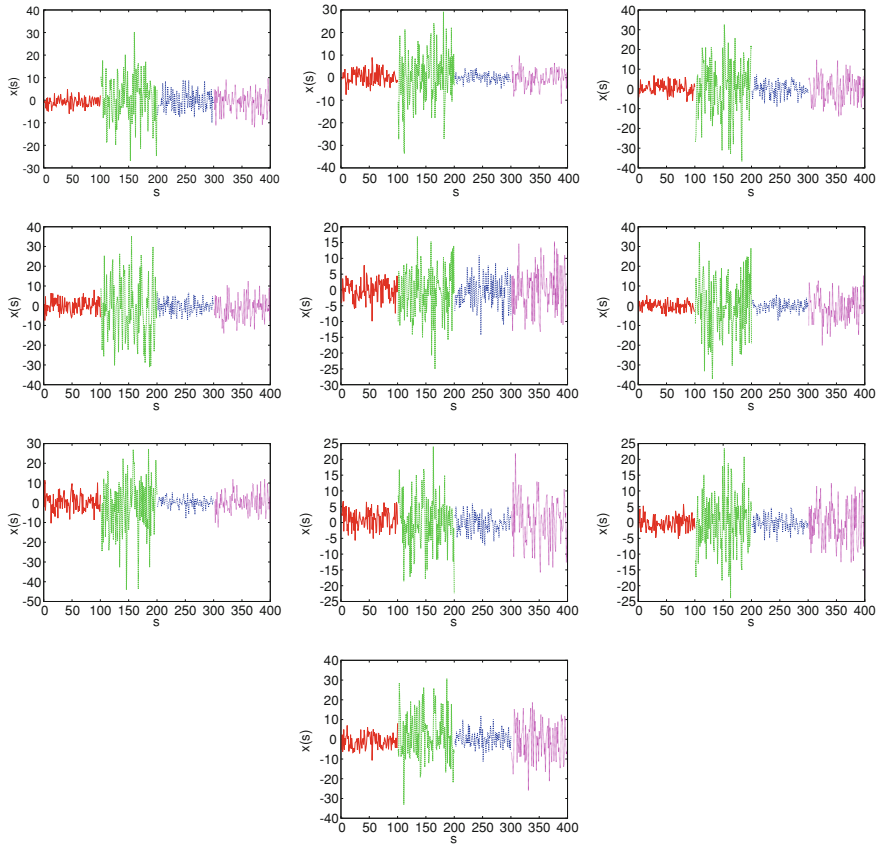


Fig. 6.1 An example of the segmented multi-dimensional time series artificially generated for $M = 10$, $T = 400$, and $\alpha_{th} = 0.01$

where $[t_{k-1}; t_k - 1]$ represents a range of the k-th segment ($t_0 = 1$, $t_1 = 101$, $t_2 = 201$, $t_3 = 301$ and $t_4 = 400$).

Setting the dimension M as 10, the length T as 400, and each segment length as 100, I generate multivariate time series and apply the proposed method to separate the time series with $\gamma = 0.3, \alpha_{th} = 0.01$ and $K = 1{,}000$ restricting that the length of each segment must be greater than $3M$. Namely, $t_{min} = 3M + 1$ and $t_{max} = T - 3M - 1$. Figure 6.1 shows an example of the segmented artificial time series. Each colour represents every segment. Four segments can be detected at exact boundaries by using the proposed method.

Table 6.1 Alphabetic codes of currencies based on ISO 4217

Abbreviation	Currency names
AUD	Australian dollar
BRL	Brazilian real
CAD	Canadian dollar
CHF	Swiss Franc
EUR	Euro
GBP	UK sterling
JPY	Japanese yen
MXN	Mexican peso
NZD	New Zealand dollar
SGD	Singapore dollar
USD	US dollar
ZAR	South African rand

6.7 Data and Empirical Analysis

In the empirical analysis, I use daily log-returns of exchange rates for 30 currency pairs[1] consisting of AUD, BRL, CAD, CHF, EUR, GBP, JPY, MXN, NZD, SGD, USD, and ZAR during the period from January 3, 2001 to December 30, 2011. Table 6.1 shows three letter alphabetic codes of currencies based on ISO 4217.

Let $R_i(s)$ be the daily exchange rate of currency pair i at time s and $r_i(s) = \ln R_i(s+1) - \ln R_i(s)$ be its daily log-return.

There are 2,760 data points in the multiple time series. The proposed segmentation procedure is applied to the process of separating the multiple log-return time series. We have 11 segments at $\alpha_{th} = 0.01$ with the restriction that the length of each segment must be greater than $3M$. Namely, $t_{min} = 3M + 1$ and $t_{max} = T - 3M - 1$. Figure 6.2 shows segmented time series for 30 currency pairs. Each colour represents a segment. The log-return time series for currency pairs shows clustered volatility and sometimes synchronously fluctuate in time.

Table 6.2 shows the period of each segment. The 6th segment corresponds to the after-shock of BNP Paribas shock in 2007, the 7th segment to after-shock of Lehman shock in 2008, and the 11th segment is related to after-shock of Euro debut crisis in 2011.

[1] The selected currency pairs are listed as AUD/JPY, BRL/JPY, CAD/JPY, CHF/JPY, EUR/AUD, EUR/BRL, EUR/CAD, EUR/CHF, EUR/GBP, EUR/JPY, EUR/MXN, EUR/NZD, EUR/SGD, EUR/USD, EUR/ZAR, GBP/JPY, MXN/JPY, NZD/JPY, SGD/JPY, USD/AUD, USD/BRL, USD/CAD, USD/CHF, USD/GBP, USD/JPY, USD/MXN, USD/NZD, USD/SGD, USD/ZAR, and ZAR/JPY.

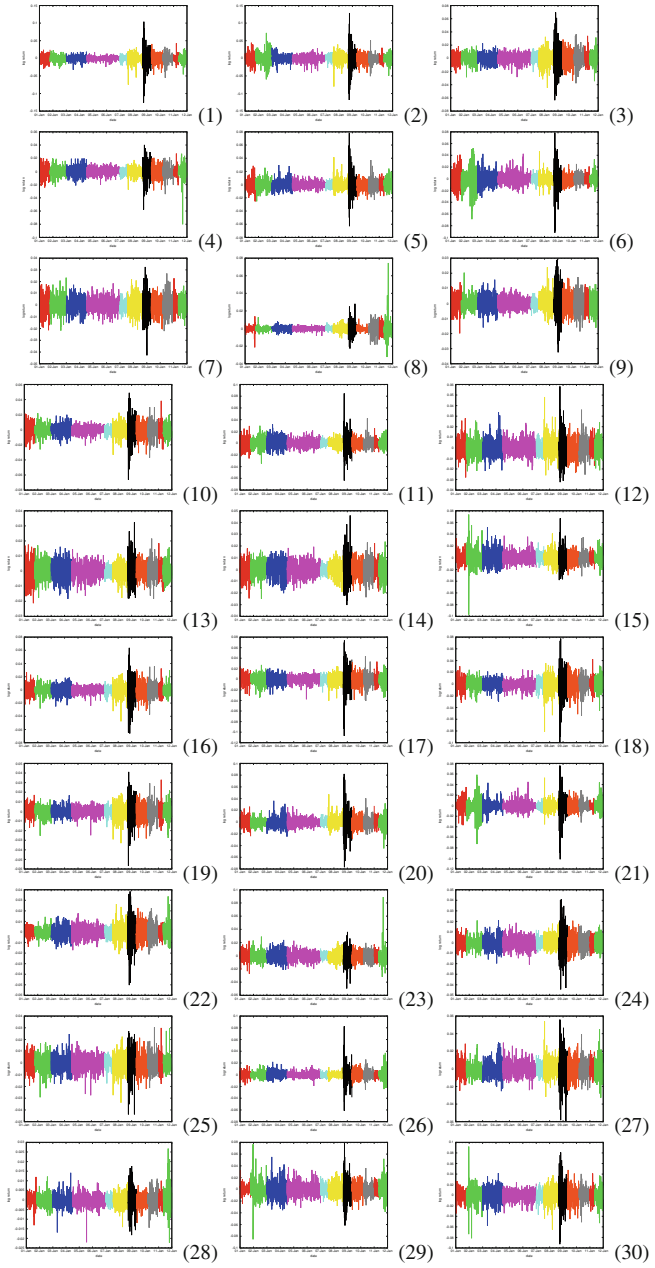


Fig. 6.2 The log-return time series of 30 currency pairs. They are separated into 11 segments. (*1*) AUD/JPY, (*2*) BRL/JPY, (*3*) CAD/JPY, (*4*) CHF/JPY, (*5*) EUR/AUD, (*6*) EUR/BRL, (*7*) EUR/CAD, (*8*) EUR/CHF, (*9*) EUR/GBP, (*10*) EUR/JPY, (*11*) EUR/MXN, (*12*) EUR/NZD, (*13*) EUR/SGD, (*14*) EUR/USD, (*15*) EUR/ZAR, (*16*) GBP/JPY, (*17*) MXN/JPY (*18*) NZD/JPY, (*19*) SGD/JPY, (*20*) USD/AUD, (*21*) USD/BRL, (*22*) USD/CAD, (*23*) USD/CHF, (*24*) USD/GBP, (*25*) USD/JPY, (*26*) USD/MXN, (*27*) USD/NZD, (*28*) USD/SGD (*29*) USD/ZAR, and (*30*) ZAR/JPY

Table 6.2 The period of each segments determined by the proposed method

k	Start date	End date
1	2001-01-03	2001-10-15
2	2001-10-16	2002-12-31
3	2003-01-02	2004-07-07
4	2004-07-08	2007-01-02
5	2007-01-03	2007-07-26
6	2007-07-27	2008-09-11
7	2008-09-12	2009-05-04
8	2009-05-05	2010-03-12
9	2010-03-15	2010-12-31
10	2011-01-04	2011-05-10
11	2011-05-11	2011-12-30

6.8 Conclusion

The information criterion (AIC) test for a mixture of multivariate Gaussian distribution was proposed. I also proposed to adopt the Jackknife method in order to evaluate statistical significance level of separation. I performed the proposed method for artificial 10-dimensional multivariate time series consisting of two segments sampled from different distributions. It was confirmed that the proposed method detects the segmented boundary with a 6% relative error. The proposed method is also applied for log-return time series consisting of 30 currency pairs and 11 segments are obtained. It was confirmed that some of segments correspond to critical events such as Paribas shock, Lehman shock, and Euro shock, respectively.

Acknowledgments The author would like to express his sincere gratitude to Prof. Zdzislaw Burda of Jagiellonian University for constructive comments and stimulating discussions.

Appendix A: Derivation of the Likelihood Function

Firstly, let us derive the likelihood function of the *i.i.d* M-dimensional Gaussian distribution $p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C})$. The log–likelihood value is calculated as follows:

$$\begin{aligned} \ln L_1 &= \sum_{s=1}^{T} \ln p(\boldsymbol{x}(s); \boldsymbol{\mu}, \boldsymbol{C}) \\ &= T \times \frac{1}{T} \sum_{s=1}^{T} \ln p(\boldsymbol{x}(s); \boldsymbol{\mu}, \boldsymbol{C}) \end{aligned}$$

$$\approx T \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_M p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C}) \ln p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C})$$
$$= -\frac{T}{2} \ln |\boldsymbol{C}| - \frac{TM}{2} \ln(2\pi) - \frac{TM}{2}. \tag{6.39}$$

Replacing true parameters $\boldsymbol{C}$ as its maximum likelihood estimators $\hat{\boldsymbol{C}}$, one has

$$\ln L_1 = -\frac{T}{2} \ln |\hat{\boldsymbol{C}}| - \frac{TM}{2} \ln(2\pi) - \frac{TM}{2}. \tag{6.40}$$

The log-likelihood value $\ln L_2(t)$ of the alternative model expressed in Eq. (6.15) is similarly computed as

$$\begin{aligned}
\ln L_2(t) = & \sum_{s=1}^{t} \ln p(\boldsymbol{x}(s); \boldsymbol{\mu}_L, \boldsymbol{C}_L) + \sum_{s=t+1}^{T} \ln p(\boldsymbol{x}(s); \boldsymbol{\mu}_R, \boldsymbol{C}_R) \\
\approx & \; t \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_M p(\boldsymbol{x}; \boldsymbol{\mu}_L, \boldsymbol{C}_L) \ln p(\boldsymbol{x}; \boldsymbol{\mu}_L, \boldsymbol{C}_L) \\
& + (T-t) \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_M p(\boldsymbol{x}; \boldsymbol{\mu}_R, \boldsymbol{C}_R) \ln p(\boldsymbol{x}; \boldsymbol{\mu}_R, \boldsymbol{C}_R) \\
= & -\frac{t}{2} \ln |\boldsymbol{C}_L| - \frac{tM}{2} \ln(2\pi) - \frac{tM}{2} \\
& - \frac{T-t}{2} \ln |\boldsymbol{C}_R| - \frac{(T-t)M}{2} \ln(2\pi) - \frac{(T-t)M}{2} \\
= & -\frac{t}{2} \ln |\boldsymbol{C}_L| - \frac{T-t}{2} \ln |\boldsymbol{C}_R| - \frac{TM}{2} \ln(2\pi) - \frac{TM}{2}.
\end{aligned} \tag{6.41}$$

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (eds.) Proceeding of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Alexandersson, H.: A homogeneity test applied to precipitation data. J. Climatol. **6**, 661–675 (1986)
3. Alexandersson, H., Moberg, A.: Homogenization of Swedish temperature data. Part I—homogeneity test for linear trends. Int. J. Climatol. **17**, 25–34 (1997)
4. Basseville, M., Nikiforov, I.V.: Detection of Abrupt Changes—Theory and Application. Prentice-Hall, Upper Saddle Revier (1993)
5. Bouchaud, J.P., Potters, M.: Theory of Financial Risks and Derivative Pricing- From Statistical Physics to Risk Management. Cambridge University Press, Cambridge (2003)
6. Brodsky, B.E., Darkhovsky, B.S.: Nonparametric Methods in Change-Point Problems. Kluwer Academic Publishers, Dordrecht (1993)

7. Burda, Z., Görlich, A., Jarosz, A., Jurkiewicz, J.: Signal and Noise in Correlation Matrix. Physica A **343**, 295–310 (2004)
8. Chen, J., Gupta, A.K.: Parametric Statistical Change Point Analysis- With Applications to Genetics. Medicine and Finance. Birkhäuser, Boston (2000)
9. Chen, J., Gupta, A.K.: Statistical inference of covariance change points in Gaussian model. Statistics **38**, 17–28 (2004)
10. Cheong, S.A., Fornia, R.P., Lee, G.H.T., Kok, J.L., Yim, W.S., Xu, D.Y., Zhang, Y.: The Japanese economy in crises—a time series segmentation study. Econ. E-J., 2012–5 (2012) URL http://www.economics-ejournal.org
11. Cho, H., Fryzlewicz, P.: Multiscale interpretation of taut string estimation and its connection to unbalanced Haar wavelets. Stat. Comput. **21**, 671–681 (2011)
12. Chow, G.C.: Tests of equality between sets of coefficients in two linear regressions. Econometrica **28**, 591–605 (1960)
13. Csörgö, M., Horváth, L.: Limit Theorems in Change-Point Analysis. Wiley, New York (1997)
14. Ducré-Robitaille, J.F., Vincent, L.A., Boulet, G.: Comparison of techniques for detection of discontinuities in temperature series. Int. J. Climatol. **23**, 1087–1101 (2003)
15. Easterling, D.R., Peterson, T.C.: A new method for detecting undocumented discontinuities in climatological time series. Int. J. Climatol. **15**, 369–377 (1995)
16. Giraitis, L., Leipus, R.: Testing and estimating in the change-point problem of the spectral function. Lith. Math. J. **32**, 15–29 (1992)
17. Giraitis, L., Leipus, R.: Functional CLT for nonparametric estimates of the spectrum and change-point problem for a spectral function. Lith. Math. J. **30**, 302–322 (1990)
18. Goldfeld, S.M., Quandt, R.E.: A Markov model for switching regressions. J. Econometrics **1**, 3–15 (1973)
19. Gopikrishnan, P., Plerou, V., Liu, Y., Amaral, L.A.N., Gabaix, X., Stanley, H.E.: Scaling and correlation in financial time series. Phys. A **287**, 362–373 (2000)
20. Gullett, D.W., Vincent, L., Sajecki, P.J.F.: Testing homogeneity in temperature series at Canadian climate stations. CCC report 90–4, Climate Research Branch, Meteorological Service of Canada, Ontario, Canada (1990).
21. Hamilton, J.D.: Regime-switching models (2005) URL dss.ucsd.edu/ jhamilto/palgrav1.pdf.
22. Hansen, B.E.: Testing for parameter instability in linear models. J. Policy Model. **14**, 517–533 (1992)
23. Hawkins, D.M.: Testing a sequence of observations for a shift in location. J. Am. Stat. Assoc. **72**, 180–186 (1977)
24. Hawkins, D.M.: Fitting multiple change-point models to data. Comput. Stat. Data Anal. **37**, 323–341 (2001)
25. Karl, T.R., Williams, C.N. Jr: An approach to adjusting climatological time series for discontinuous inhomogeneities. J. Clim. Appl. Meteorol. **26**, 1744–1763 (1987)
26. Kawahara, Y., Sugiyama, M.: Sequential change-point detection basedon direct density-ratio estimation. Stat. Anal. Data Min. **5**, 114–127 (2012)
27. Kim, C.J., Piger, J.M., Startz, R.: Estimation of Markov Regime-switching regression Models with endogenous swithing. Working Paper 2003–015C, Federal Researve Bank of St. Luis (2003) URL http://research.stlouisfed.org/wp/2003/2003-015.pdf
28. Laloux, L., Cizeau, P., Bouchaud, J.P., Potters, M.: Noise dressing of financial correlation matrices. Phys. Rev. Lett. **83**, 1467–1470 (1999)
29. Lavielle, M., Teyssière, G.: Detection of multiple change-points in multivariate time series. Lith. Math. J. **46**, 287–306 (2006)
30. Mandelbrot, B.: The variation of certain speculative prices. J. Bus. **36**, 394–419 (1963)
31. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics- Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2000)
32. Miao, B.Q., Zhao, L.C.: Detection of change points using rank methods. Commun. Stat. **17**, 3207–3217 (1988)
33. Neyman, J., Pearson, K.: On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. Lond. A **231**, 289–337 (1933)

34. Ombao, H., Von Sachs, R., Guo, W.: SLEX analysis of multivariate nonstationary time series. J. Am. Stat. Assoc. **100**, 519–531 (2005)
35. Ombao, H.C., Raz, J.A., Von Sachs, R., Guo, W.: The SLEX model of a non-stationary random process. Ann. Inst. Stat. Math. **54**, 171–200 (2002)
36. Papp, G., Pafka, S., Nowak, M.A., Kondor, I.: Random matrix filtering in portfolio optimization. Acta Phys. Pol. B **36**, 2757–2765 (2005)
37. Perreault, L., Haché, M., Slivitzky, M., Bobée, B.: Detection of changes in precipitation and runoff over eastern Canada and US using a Bayesian approach. Stochast. Environ. Res. Risk Assess. **13**, 201–216 (1999)
38. Perreault, L., Bernier, J., Bobée, B., Parent, E.: Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. J. Hydrol. **235**, 221–241 (2000)
39. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: Universal and nonuniversal properties of cross correlations in financial time series. Phys. Rev. Lett. **83**, 1471–1474 (1999)
40. Quandt, R.E., Ramsey, J.B.: Estimating mixtures of normal distributions and switching regressions. J. Am. Stat. Assoc. **73**, 730–738 (1978)
41. Sato, A.-H.: Recursive segmentation procedure based on the Akaike information criterion test. 2013 IEEE 37th Annual Signature Conference of Computer Software and Applications Conference (COMPSAC), pp. 226–233 (2013)
42. Sen, A., Srivastava, M.S.: On tests for detecting change in the mean. The Annals of Statistics **3**, 98–108 (1975)
43. Shiryaev, A.N., Zhitlukhin, M.V.: Optimal stopping problems for a Brownian motion with a disorder on a finite interval (2012) arXiv:1212.3709
44. Vert, J., Bleakley, K.: Fast detection of multiple change-points shared by many signals using group LARS. Adv. Neural Info. Process. Syst. **23**, 2343–2351 (2010)
45. Vincent, L.A.: A technique for the identification of inhomogeneities in Canadian temperature series. J. Clim. **11**, 1094–1104 (1998)
46. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)
47. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B **68**, 49–67 (2006)

Chapter 7
Hotel Booking Data

Abstract This study considers a method to determine and classify districts based on the stay capacity of hotels in order to understand regional dependence of social wealth. We analyse the geographical positions and the number of rooms about 2,881 Japanese hotels which have 582,898 rooms in total empirically. Firstly, we conduct a clustering analysis of regional statistics on the stay capacities by using the centroid method. Secondly, we divide areas by a centroid method from a maximum entropy point of view hierarchically. It may be concluded that the rank size distribution for the number of rooms in the cluster is fitted with a power-law function with the exponent depending on the number of clusters included in the level. We further investigates an association between the availability of hotels and socioeconomic dynamics before and after the Great East Japan Earthquake on 11 March, 2011.

7.1 Introduction

The regional statistics provide useful quantitative methods for understanding situations of socioeconomic systems [7]. One direction of research purposes is to measure societal stocks. For example, we may characterise socioeconomic states based on regional dependence of social wealth. The other is to characterise societal flows such as migration, monetary flow, and logistics. These flows are deeply related to the regional dependence of the social stocks.

It is known that the wealth distribution is fitted with the power-law function, referred to as Pareto distributions. A size distribution of populations in cities is also known as the power-law distribution. Since these power-law properties are believed to be generated through preferential attachment mechanism [1], we may also find some relationship between regional dependence of socioeconomic activities and the power-law distributions.

The migration processes have been intensively studies in the context of socioeconomic dynamics with particular interests for quantitative research [20]. Weidlich and Haag proposed the Master equation with transition probabilities

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_7, © Springer Japan 2014

depending on both regional-dependent and time-dependent utility and mobility in order to describe collective tendency of agent decision in migration chance [9].

The tourism industry obtains profits from demands of temporal migration. Therefore, what we examine stay capacities of hotels included in areas may provide insights on relationship between the social wealth and the migration process. In this chapter, we investigate regional dependence of social wealth based on the data on Japanese hotel industry with geographical information. By using data on room capacities as proxy variables of the regional dependence of wealth distribution, we propose a method to characterise a spatial density of Japanese economy.

In Japan, there are over 54,000 accommodations [11], which are rich in various types: from the largest hotel with over 3,000 rooms to the highest class Japanese inn with few rooms. Their types and capacities also depend on a district.

According to the study of tourism management [3], there are push and pull factors, so that tourism motivation is determined by the situation of the travellers (push) and the situation of the destination (pull). The idea behind this two-dimensional approach is that people travel because they are pushed by their own internal forces and pulled by the external forces of the destination attributes [6]. The pull factors originate from the destination properties (supply). More recently, Tkaczynski et al. applied the stake-holder theory, a management theory proposed by Freeman [8], to a destination in tourism [19]. The existence of hotel accommodations implies that pull factors are present in the district where they are located. In the context of economics, this means that the demand-supply situation is generated by both consumers and suppliers. Namely, they can be dependent on the area and the season [5].

Moreover, a problem for estimating demand from censored booking data has been recognised for many years in the hotel industry. Patrick et al. [12] developed parametric regression models that consider not only the demand distribution, but also the conditions under which the data were collected. Sato [13] investigated regional patterns of Japanese travel behaviour by using the EM algorithm for finite mixtures of Poisson distributions. Aftereffects of some events can be observed from activities of tourism industry [14].

Therefore, we may assume that demand and supply in the hotel industry can reflect both the social and economic situations. In this chapter, we collect Japanese hotel data from Jalan [10], which is one of the most famous hotel booking sites in Japan. We analyse regional hotel distribution from this. Particularly, we define areas in terms of their hotel capacities with a hierarchical classification method. Recently, Chen proposed the maximum entropy principle on the city size distribution [4]. This concept is applicable to defining the regional category for our purpose. Using the maximum entropy principle on the total number of rooms in the areas, we propose a method to decide hierarchical structure of capacity.

This chapter is organised as follows. In Sect. 7.2, the data description is briefly presented. In Sect. 7.3, characteristics of data on hotel opportunities are shown. In Sect. 7.4, we determine regional dependence on the number of rooms and classify its districts. In Sect. 7.5, we examine relationship between the hotel availability and physical damage of the Great East Japan Earthquake on 11 March, 2011. Finally, Sect. 7.6 is devoted to concluding remarks.

7.2 Data Description

In this section, we give a brief explanation of our data. We used data collected from a Japanese hotel booking site named Jalan. The data contains hotel identifiers ("hotelid"), geographical position ("latitude" and "longitude"), area identifiers ("prefecture", "large area" and "small area"), and room capacity ("number of rooms"). The area identifiers ("large area" and "small area") are defined by Jalan.

Firstly, we explain the Jalan Web Application Programming Interface (API) to collect data on accommodations. The Jalan Web API is a source code intended to be used as an interface by software components to communicate with the Jalan Web server. Third party can build a new web application service with the Jalan Web API. Figure 7.1 shows a conceptual illustration of the Jalan web service. The Jalan server can store various information on accommodations. The hotel managers input basic information about their accommodation into the Jalan server via a Web interface. The customers can search their preference from available information stored in the Jalan server via their Internet browsers.

The data contains latitude, longitude, prefecture, large area, small area, and hotelid, but does not contain the number of rooms. Moreover, we accessed hotel homepages on the Jalan web site and got the HTML document including the number of rooms of every hotel. In order to get each hotel capacity automatically, we have developed a HTML parser which extracts the portion describing its own room capacity from the document. After extracting the capacities, we linked them with the hotel locations by using the hotel identifiers.

7.3 Outlook

In the data set, there exist over 100,000 room opportunities at over 14,000 hotels. Table 7.1 shows contents included in the data set. Each plan contains sampled date, stay date, regional sequential number, hotel identification number, hotel name, postal address, URL of the hotel website, geographical position, plan name, and rate.

Since the data contains regional information, it is possible for us to analyse regional dependence of hotel rates. Throughout the investigation, we regard the number of recorded opportunities (plan) as a proxy variable of the number of available room stocks.

First of all, we show the data for the period from 24 December, 2009 to 8 May, 2011. The data is missing from 14 to 30 March, 2011 because the the web service was not available due to the Great East Earthquake. Figure 7.2 shows an example of distributions and representative rates. An example of rates distributions under the condition that two adults can stay at the hotel for one night at 23 December, 2009. This data have been sampled on 25 December, 2009. The yellow to black filled squares represent hotel plans costing ranging from 1,000 JPY and 50,000 JPY per

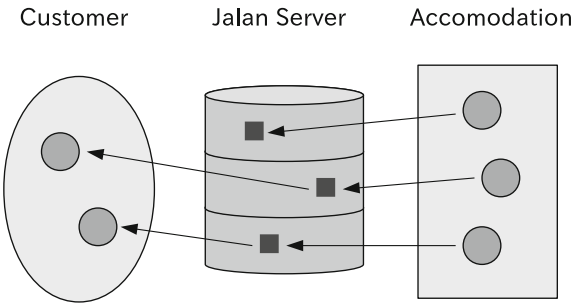


Fig. 7.1 A conceptual illustration of the Jalan web service. The hotel information is input on the Jalan server. Customers can browse the basic information such as a number of rooms via the Jalan web page

Table 7.1 The data format of room opportunities

Date of collection
Date of stay
Hotel identification number
Hotel name
Hotel name (kana characters)
Postal code
Address
URL
Latitude
Longitude
Opportunity name
Meal availability
The latest best rate per night
Rate per night

night. The red filled squares represent hotel plans costing over 50,000 JPY per night. We found that there was a strong dependence of vacancies on places. Specifically, we find that many hotels are located around several centralised cities such as Tokyo, Osaka, Nagoya, Fukuoka, and so on.

The number of one-night, twin-share room plans was counted from the recorded csv files throughout the whole sampled period. Figure 7.3 shows the daily number of room opportunities with different durations D, which is defined as a difference between stay date and sampling date. From this graph, we found three facts:

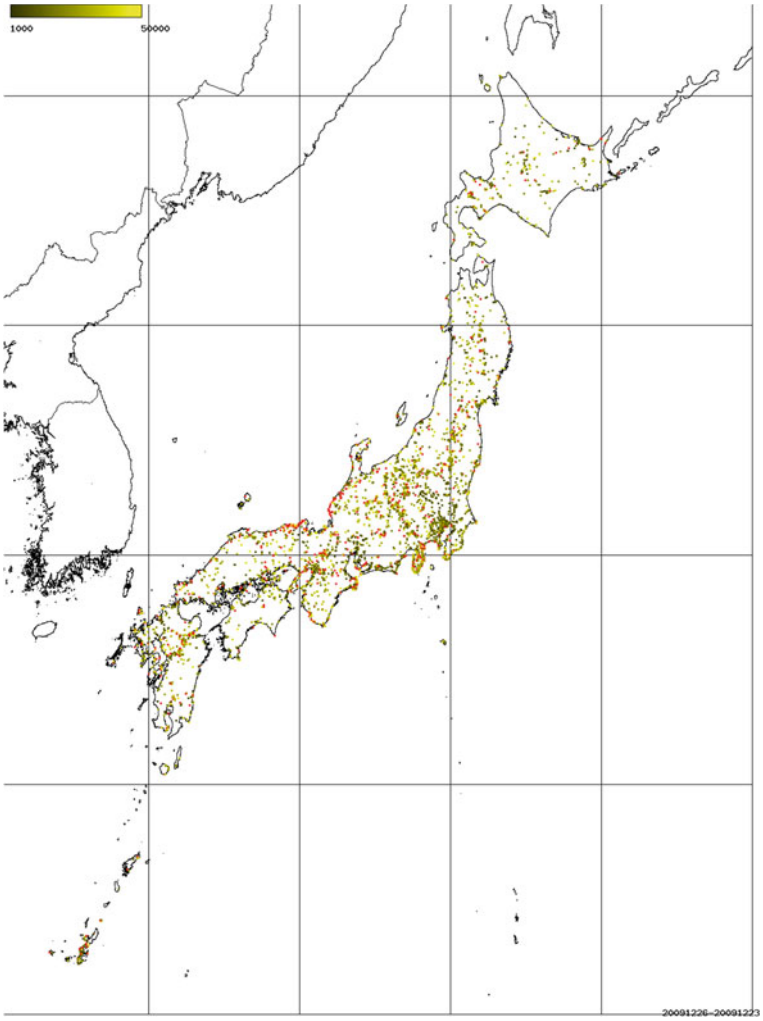


Fig. 7.2 The regional dependence of room prices of Japanese available on 26 December, 2009, as of 23 December, 2009

(1) The number of room opportunities shows a weekly seasonality.
(2) There is a strong dependence of the number of available opportunities on the Japanese calendar. Namely, Saturdays and holidays drove reservation activities of consumers. For example, during the New Year holidays (around from 30 December to 3 January) and holidays in the spring season (around 20 March), the time series of the numbers show big drops.

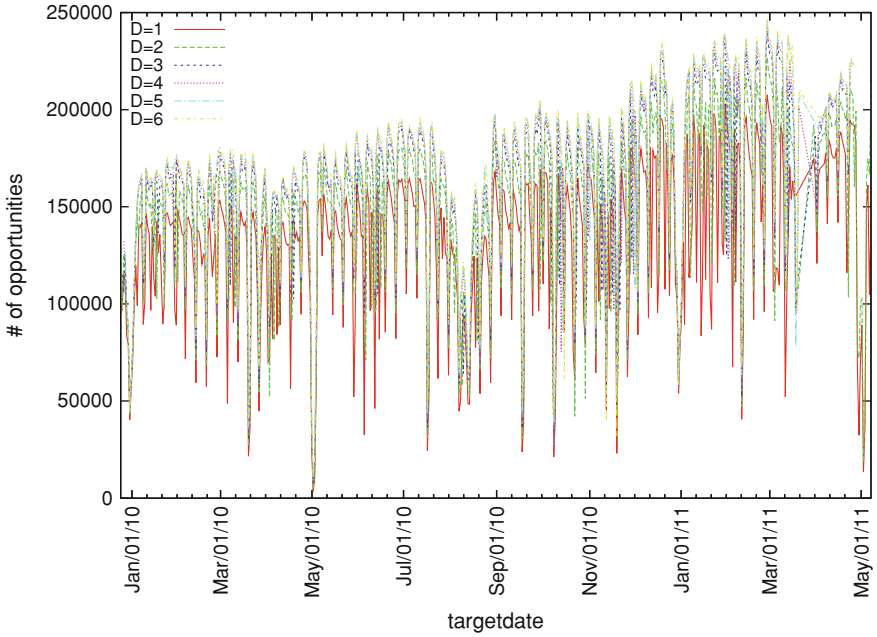


Fig. 7.3 The number of room plans for one-night stay by two adults during the period between 24 December, 2009 to 8 May, 2011

(3) The number eventually increases as the date of stay reaches. Specifically, it is observed that the number of opportunities drastically decreases two days before the date of stay.

Furthermore, Figure 7.4 shows dependence of average rates all over the Japan on calendar dates with different durations. During the New Year holidays in 2010, the average rates rapidly decreased. Meanwhile, on the spring holidays in 2010, the average rates rapidly increased. This difference seems to arise from the difference of consumers' motivation structure and preference on price levels between these holiday seasons.

Figure 7.5 shows scatter plots between the daily number of room opportunities and average of room rates. The high-demand dates exhibit larger variations of the average rate than low-demand dates. The preferable price level of consumers has a high variability on high-demand dates.

7.4 Hotel Rank Distribution

Here, we focus on accommodations which possess more than 100 rooms. Figure 7.6 represents geographical position of accommodations on a map of Japan on January 11, 2012. As shown in Fig. 7.6, the large cities, such as Tokyo and Osaka, there are

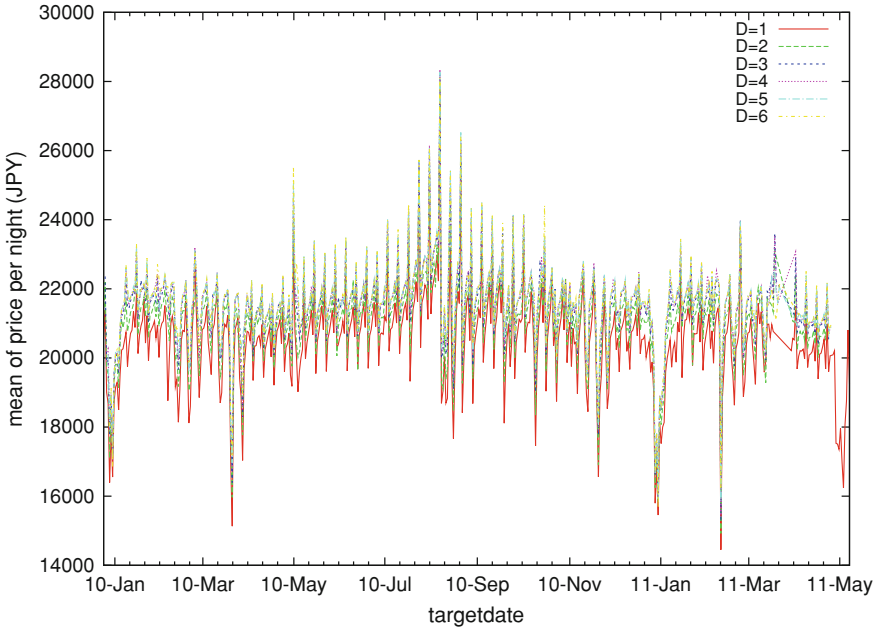


Fig. 7.4 Average prices for one-night, twin-share room plans throughout Japan for the period from 24 December, 2009 to 8 May, 2011

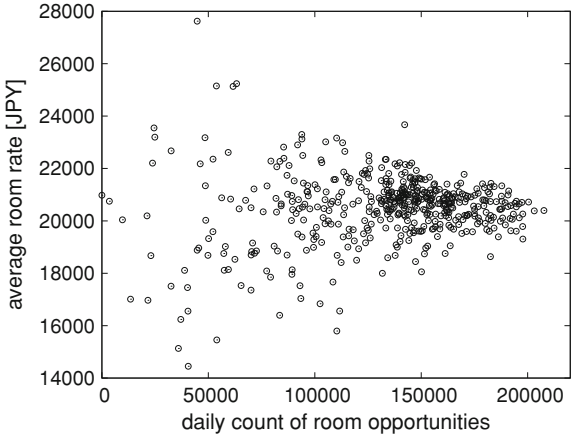


Fig. 7.5 Scatter plots between the daily number of room opportunities and mean room rates across Japan for the period from 24 December, 2009 to 10 June, 2011

more hotels than other local cities. Obviously, this fact suggests that the larger cities have the larger stay capacities. Table 7.2 shows the total number of hotels, of rooms, of large areas and of small areas. The 582,898 rooms in 2,881 accommodations are

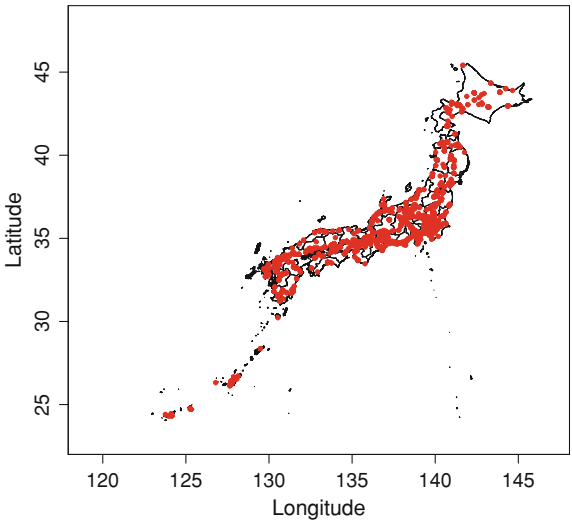


Fig. 7.6 The position of accommodations with more than 100 rooms on a Japanese map on January 11, 2012. The x-axis represents longitude, and the y-axis latitude

Table 7.2 Summary of accommodations with more than 100 rooms

# Hotels	# Rooms	# Large area	# Small area
2,881	582,898	311	682

distributed in 311 large areas consisting of 682 small areas. According to the Japan Tourism Agency of the Ministry of Land, Infrastructure, Transport and Tourism [11], it is reported that there exist 6,390 hotels where more than 10 employees work in Japan. Assuming that there are 10 or more employees in every hotel which has more than 100 rooms, we estimate coverage of hotels which are available in Jalan as 45.1 %.

We sort hotels in descending order about the number of rooms and examine a relation between the room capacity of each accommodation and its rank. This relationship is generally called rank size distribution. Figure 7.7 shows that the rank size distribution is approximated as a power-law function

$$R_k = R_1 k^{-q}, \tag{7.1}$$

where R_k shows the number of rooms at the rank k hotel, q represents a scaling exponent ($q > 0$) and R_1 is a positive constant ($R_1 = 2829.3673$ and $q = 0.3917$). The power law rank distribution is often observed in natural and economic phenomena.

This figure shows that there exist hotels more than 1,000 rooms from the first to 10th largest. We found that accommodations from 100 rooms to 500 rooms are ranked from the top 100 to the top 1,000. We calculate $q \simeq 0.3917$ by using the OLS regression.

Figure 7.8 shows geographical representation of a stay capacity at each hotel. The diameters represent the stay capacities, and the centre coordinates are geographical positions. It seems that in the capital city Tokyo, a large number of rooms is accumulated.

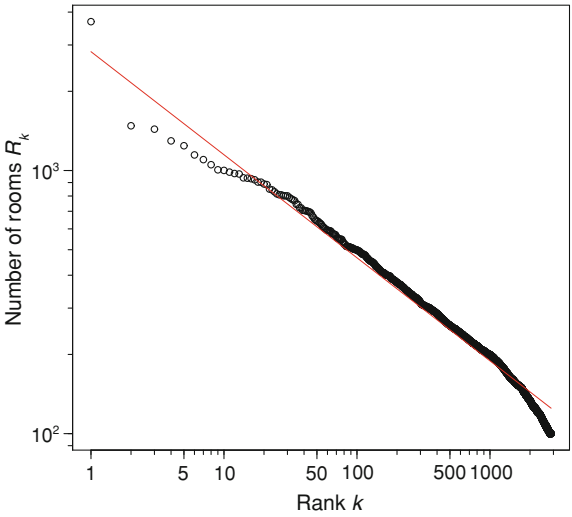


Fig. 7.7 The number of rooms R_k as a function of its rank k in a double logarithmic scale. The *solid line* represents the power-law relationship computed with the OLS regression: $R_k = 2829.3673k^{-0.3917}$

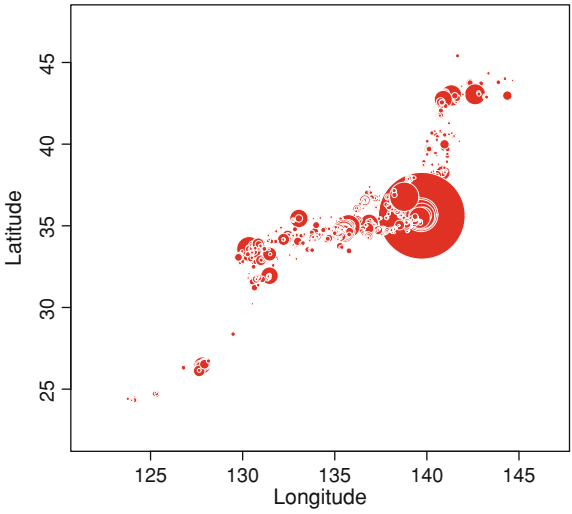


Fig. 7.8 The geographical position and capacity of every hotel. The x-axis is longitude, and the y-axis latitude. The centre coordinates of *circle* are the longitude and latitude of the hotel. The *diameter* represents the number of rooms of each hotel

7.4.1 Method

In this section, we explain a clustering method to define districts in terms of stay capacity and their hierarchical structure based on the maximum entropy principle. Because we want to measure regional stay capacity, we use a clustering with the centroid method. Because we want to know regional level, we define hierarchical structure with the maximum entropy principle.

7.4.1.1 Clustering by Centroid Method

We give an explanation of the clustering algorithm used throughout the investigation in this subsection.

Let e_i $(i = 1, \ldots, T)$ represent the ith hotel identifier consisting of its longitude x_i, latitude y_i and capacity c_i, where T is the total number of hotels. Let $\mathscr{H}$ further denote a set $\{\mathscr{H}_0, \ldots, \mathscr{H}_{T-1}\}$ of nested clustering, where $\mathscr{H}_t (t = 0, \ldots, T-1)$ is a clustering at step t. Suppose that $E_j (j = 1, \ldots, T)$ is the jth cluster, which is initially $E_j = \{e_j\}$, $(j = 1, \ldots, T)$. Let $d(E_i, E_j)$ be a distance function between clusters E_i and E_j. Then, the clustering algorithm is described as shown in Algorithm 1.

Algorithm 1 Clustering by centroid method

Require: A set $\{e_1, \ldots, e_T\}$ of data points
Ensure: A set $\mathscr{H} = \{\mathscr{H}_0, \ldots, \mathscr{H}_{T-1}\}$ of nested clustering
  begin
  Initialisation: $\mathscr{H}_0 = \{E_1, \ldots, E_T\}$ where $E_k = \{\mathbf{e}_k\}$
  repeat
    Selection: $_{E_i, E_j} d(E_i, E_j)$
    Merge: $\mathscr{H}_{t+1} = (\mathscr{H}_t \backslash \{E_i, E_j\}) \cup \{E_i \cup E_j\}$
  until all points belong to the same cluster
  end

The clustering method consists of three phases: Initialisation, Selection, and Merge. After the Initialisation phase, the clustering algorithm repeats the Selection and Merge phases until all points belong to the same cluster. Note that this iteration can be stopped at any step t.

The Selection phase works based on the $T \times T$ distance matrix that requires $O(T^2)$ space and time complexities. In the selection phase, the distance function between clusters is defined as

$$d(E_i, E_j) = \sqrt{(x_i^G - x_j^G)^2 + (y_i^G - y_j^G)^2}, \tag{7.2}$$

where (x_i^G, y_i^G) and (x_j^G, y_j^G) are centroids of E_i and E_j, respectively. A coordination of the centroid of E_i is expressed as

$$x_i^G = \frac{\sum_{e_{i'} \in E_i} c_{i'} x_{i'}}{\sum_{e_{i'} \in E_i} c_{i'}}, \tag{7.3}$$

$$y_i^G = \frac{\sum_{e_{i'} \in E_i} c_{i'} y_{i'}}{\sum_{e_{i'} \in E_i} c_{i'}}. \tag{7.4}$$

The Merge phase removes the nearest pair of clusters, E_i and E_j, from $\mathscr{H}_t$ and then adds a union of these two clusters in order to formulate the cluster $\mathscr{H}_{t+1}$ at the next step. A union of these two clusters has elements $(x_{i^*}, y_{i^*}, c_{i^*})$, given by

$$x_{i^*} = \frac{c_i x_i + c_j x_j}{c_i + c_j}, \tag{7.5}$$

$$y_{i^*} = \frac{c_i y_i + c_j y_j}{c_i + c_j}, \tag{7.6}$$

$$c_{i^*} = c_i + c_j. \tag{7.7}$$

If two clusters are merged into a single cluster at a certain step, then they will remain in the same cluster for all subsequent clustering. Because there are $T - t$ clusters at step t, there are $\frac{(T-t)(T-t-1)}{2}$ pairs to compare to find the nearest pair at that step. Note that the algorithm needs to repeat up to $T - 1$ times; thus, this clustering method requires $O(T^3)$ time complexity.

7.4.1.2 Maximum Entropy Principle

The general form of the rank size scaling law can be expressed as [4]

$$R_k = R_1 k^{-q}, \tag{7.8}$$

where k denotes the rank by the capacity of the areas, R_k refers to the number of rooms of the kth area, R_1 to the number of rooms of the largest area, and q, the scaling exponent of the rank size distribution.

Suppose that there is a region $\boldsymbol{R}$ consisting of n number of subareas and that there are N rooms within the region $\boldsymbol{R}$. Here, we consider that we classify the areas into M levels and form a hierarchy.

7.4.1.3 Total Room Number

We maximise the entropy of the total number of room capacities. Let f_m and C_m be the number of districts at the mth level and the mean size over the f_m districts, respectively. The number of room capacities at the mth level, denoted as S_m, may be described as

$$f_m C_m = S_m, \quad m = 1, \ldots, M. \tag{7.9}$$

The state number of the N rooms in different M classes, W_S, can be expressed as a problem of ordered partition of the room set. In fact, an ordered partition of "typeS_1 + $\cdots$ + typeS_M" is one in which the mth part has S_m rooms, for $m = 1, \ldots, M$. The state number of such partitions is given by the following multinomial coefficient:

$$W_S = \binom{N}{S_1, \ldots, S_M} = \frac{N!}{S_1! \cdots S_M!}. \tag{7.10}$$

Thus, the information entropy function is defined as

$$\begin{aligned} H_S &= \ln W_s \\ &= \ln N! - \sum_{m=1}^{M} \ln S_m!. \end{aligned} \tag{7.11}$$

Regarding that the total number of room capacities N is constant, we may describe the maximum entropy problem as

$$\{\hat{s}_1, \ldots, \hat{s}_m\} = \arg\max \ H_S, \tag{7.12}$$

$$\text{s.t.} \sum_{m=1}^{M} \frac{S_m}{N} = 1. \tag{7.13}$$

Equation (7.13) means that the entropy is maximised on the condition where the summation of room capacities over different classes equals N. If M is finite, then a Lagrange function of the above nonlinear programming problem can be defined by

$$L_S = \ln N! - \sum_{m=1}^{M} \ln S_m! + \lambda \left(\sum_{m=1}^{M} S_m - N \right), \tag{7.14}$$

where λ is a Lagrange multiplier. According to the condition of extreme value, derivative of $L(S)$ with respect to S_m $(m = 1, \ldots, M)$ yields

$$S_m = e^{\lambda} = Const., \tag{7.15}$$

where e^{λ} is a positive constant and S_m is independent of m. Inserting Eq. (7.13) into Eq. (7.15), we have $S_m = N/M$ $(m = 1, \ldots, M)$. Equations (7.9) and (7.15) imply the relation

$$f_m = \eta C_m^{-1}, \tag{7.16}$$

where we set $\eta = e^{\lambda}$.

7.4.1.4 The Number of Districts and Averaged Number of Room Capacities

The state number of n districts included in different M classes, W_f, can be expressed as a problem of ordered partition of the district set. In fact, an ordered partition of "type $f_1 + \cdots +$ type f_M" is one in which the mth part has f_m members, for

$m = 1, \ldots, M$. The state number of such partitions is given by the following multinomial coefficient:

$$W_f = \binom{n}{f_1, \ldots, f_M} = \frac{n!}{f_1! \cdots f_M!}, \tag{7.17}$$

where $m = 1, \ldots, M$ denotes the ordinal number of district levels in hierarchy. Thus, the information entropy of frequency distribution of districts is described as

$$H_f \equiv \ln W_f = \ln n! - \sum_{m=1}^{M} \ln f_m!, \tag{7.18}$$

where H_f refers to the information entropy of frequency distribution.

Let K represent the summation of the averaged number of room capacities in different classes,

$$K = \sum_{m=1}^{M} C_m. \tag{7.19}$$

The state number of the averaged number of the capacities in the hierarchy based on top-down order, W_C, can be expressed as an ordered partition problem and defined by

$$W_C = \binom{K}{C_1, \ldots, C_M} = \frac{K!}{C_1! \cdots C_M!}. \tag{7.20}$$

The information entropy of the size distribution is described as

$$H_C \equiv \ln W_C = \ln K! - \sum_{m=1}^{M} \ln C_m!, \tag{7.21}$$

where H_C refers to the information entropy of size distribution.

Here, let us assume that these entropies, H_f and H_C, are maximised at the same time. Then a nonlinear programming problem can be built as follows:

$$\text{Max } H_f + \mu H_C = \ln W_f + \mu \ln W_C, \tag{7.22}$$

$$\text{S.t. } \sum_{m=1}^{M} \frac{f_m}{n} = 1, \tag{7.23}$$

$$\sum_{m=1}^{M} m \frac{f_m}{n} = \omega, \quad (n > f_M), \tag{7.24}$$

$$\sum_{m=1}^{M} \frac{C_m}{K} = 1, \tag{7.25}$$

$$\sum_{m=1}^{M} m \frac{C_m}{K} = \varphi, \quad (K > C_m), \tag{7.26}$$

$$f_m C_m = S_m, \quad (m = 1, \ldots, M), \tag{7.27}$$

where ω and φ are positive constants, μ represents an arbitrary constant. The former implies that the mean of levels is finite. The latter implies that the mean of capacities is finite. This is a kind of optimisation problem. The first constraint condition, Eq. (7.23), indicates that the number in a district is also constant. The second constraint Eq. (7.24) indicates that there exists an expectation value of area level. The third constraint condition, Eq. (7.25), indicates that the sum of the averaged number of rooms in a district is constant. The fourth constraint Eq. (7.26) indicates that there exists an expectation value of averaged number of rooms.

In order to solve the aforementioned programming problem, we can construct a Lagrange function such as

$$\begin{aligned} L = \ln n! - \sum_{m=1}^{M} \ln f_m! + \lambda_1 \left(n - \sum_{m=1}^{M} f_m \right) + \lambda_2 \left(n\omega - \sum_{m=1}^{M} m f_m \right) \\ + \mu \ln K! - \mu \sum_{m=1}^{M} \ln C_m! + \lambda_3 \left(K - \sum_{m=1}^{M} C_m \right) + \lambda_4 \left(K\varphi - \sum_{m=1}^{M} m C_m \right) \\ + \lambda_5 (S_1 - f_1 C_1) + \cdots + \lambda_{4+M} (S_m - f_M C_M), \end{aligned} \tag{7.28}$$

where $\lambda_1, \ldots, \lambda_{4+M}$ are Lagrange multipliers. According to Stirling's formula, $\ln x! \simeq x \ln x - x$ is satisfied, for a sufficiently large integer x. Therefore, we obtain an approximate expression in $\frac{\mathrm{d} \ln x!}{\mathrm{d}x} \simeq \ln x$. Namely, if n, f_m, K and C_m are large enough, then $\frac{\mathrm{d} \ln n!}{\mathrm{d}n} \simeq \ln n$, $\frac{\mathrm{d} \ln f_m!}{\mathrm{d} f_m} \simeq \ln f_m$, $\frac{\mathrm{d} \ln K!}{\mathrm{d}x} \simeq \ln K$ and $\frac{\mathrm{d} \ln C_m!}{\mathrm{d} C_m} \simeq \ln C_m$. We consider the Lagrangian condition of extreme value:

$$\frac{\partial L}{\partial f_m} = -\ln f_m - \lambda_1 - \lambda_2 m - \lambda_{4+m} C_m = 0, \tag{7.29}$$

$$\frac{\partial L}{\partial C_m} = -\mu \ln C_m - \lambda_3 - \lambda_4 m - \lambda_{4+m} f_m = 0. \tag{7.30}$$

From Eqs. (7.29) and (7.30), respectively, we can introduce

$$f_m = f_0 \exp(-\lambda_2 m - \lambda_{4+m} C_m), \tag{7.31}$$

$$C_m = C_0 \exp\left(\frac{-\lambda_4 m - \lambda_{4+m} f_m}{\mu} \right), \tag{7.32}$$

where $f_0 = \exp(-\lambda_1)$ and $C_0 = \exp(-\lambda_3/\mu)$. Considering Eq. (7.16), we get

$$f_m C_m = f_0 C_0 \exp\left(-\lambda_2 m - \lambda_{4+m} C_m - \frac{\lambda_4 m}{\mu} - \frac{\lambda_{4+m} f_m}{\mu}\right) = \eta. \tag{7.33}$$

Taking the logarithm of Eq. (7.33), we have

$$\left(\lambda_2 + \frac{\lambda_4}{\mu}\right) m + \frac{\lambda_{4+m}}{\mu} f_m + \lambda_{4+m} C_m + \ln \frac{\eta}{f_0 C_0} = 0. \tag{7.34}$$

Since Eq. (7.34) is the identical equation in terms of m, f_m, C_m, we have the following relation:

$$\lambda_2 + \frac{\lambda_4}{\mu} = 0, \quad \lambda_{4+m} = 0, \quad \eta = f_0 C_0. \tag{7.35}$$

Hence, we can rewrite Eqs. (7.31) and (7.32) as

$$f_m = f_0 e^{-\lambda_2 m}, \tag{7.36}$$

$$C_m = C_0 e^{\lambda_2 m}. \tag{7.37}$$

The number of districts and the averaged number of room capacities, respectively, at the mth level are exponential functions in terms of m.

7.4.2 Results and Discussion

In order to estimate regional stay capacity, we use a method to cluster accommodations based on their locations and capacities. As shown in Sect. 7.4.1.1, we can arbitrarily determine the number of clusters. Here, we fix the cluster number to be 311 because Jalan identifies 311 large areas. Figure 7.9 shows the rank-size relationship. We found the power-law relationship between rank k and the number of rooms R_k. The power law exponent is estimated as 0.9787 by means of the OLS regression. Note that the scaling exponent is nearly equal to 1.

Figure 7.10 represents the relation between the number of clusters τ and the scaling exponent q, where we set $\tau = T - t$. As shown in this figure, the scaling exponent tends to decrease as the number of clusters increases. From $\tau = 219$ to $\tau = 482$, the scaling exponent exists around 1. The scaling exponent decreases rapidly when the number of clusters is 482, and then falls below 1. In addition, if the number of clusters is under 219, the exponent fluctuates steeply. The cluster number 311, which is equal to large area number, exists between the 219 clusters and the 482 clusters. Figure 7.11 shows the regional dependence of stay capacity computed with the proposed method. The diameter represents the stay capacity of the cluster, and

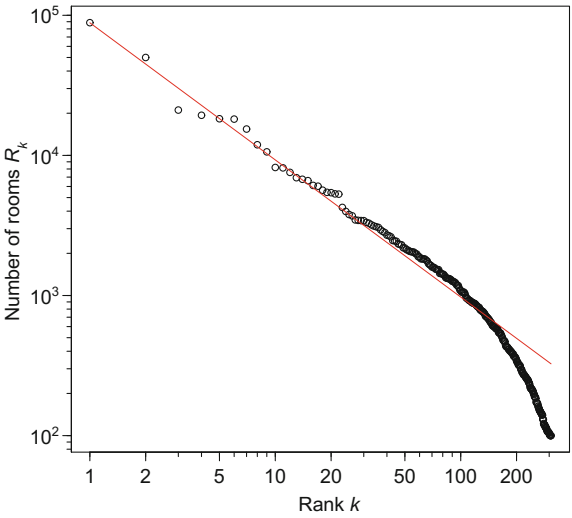


Fig. 7.9 The room number of the cluster R_k as a function of its rank k in a log–log scale. The *solid line* represents the power-law relationship computed with the OLS regression: $R_k \simeq 88416.7982k^{-0.9787}$

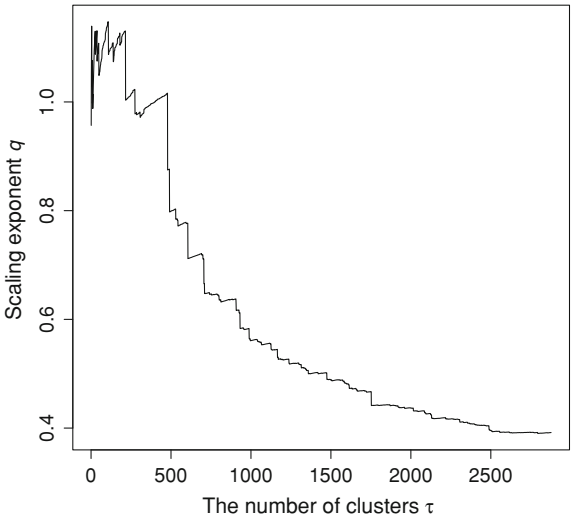


Fig. 7.10 The relation between the number of clusters τ and an power exponent q

the centre coordinates are the centroids of the cluster. Tokyo and Osaka are extracted as the largest cluster and the second largest.

Let us classify the 311 areas into 4 levels. As shown in Sect. 7.4.1.2, we can arbitrarily determine what level the areas is divided into, and the total number of rooms in each level are equal. Moreover, the number of areas in each level increases exponentially, and the average number of rooms decreases exponentially. Here, we separate every area in the descending order so that the total number of rooms in each level becomes equal. We confirmed that there is clear relationship among m, f_m and

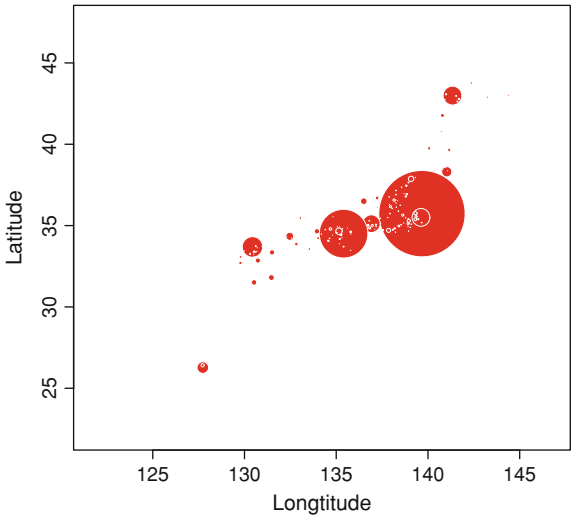


Fig. 7.11 The geographical position of every cluster and capacity. The x-axis is longitude, and the y-axis latitude. The centre coordinates of *circle* are the longitude and latitude of the cluster. The *diameter* shows the total room number of the cluster

C_m. Figures 7.12 and 7.13 show their relations. We found that f_m and C_m are fitted with an exponential function in terms of m.

The cities of level 1 are Tokyo and Osaka which are the largest city and the second largest city. The cities of level 2 are government-decreed city (see Table 7.3). The cities of level 3 are provincial central city, and those of level 4 are rest.

7.5 Impact of Natural Disasters (Great East Japan Earthquake on 11 March, 2011)

Since people are also products of nature, the physical effects of the natural environment on our society are remarkable. Specifically, natural disasters often affect our societies significantly. Therefore, we need to understand the subsequent impact of natural disasters on human behaviour, from both economical and social perspectives.

The first Great East Japan Earthquake hit at 14:46 on 11 March, 2011 in Japanese local time (05:46 in UTC). Within 20 min, huge tsunamis had devastated cities along Japan's northeastern coastline. In addition to wide-spread physical destruction, social infrastructures also suffered extensive damaged. It is important for us to understand its subsequent impact on our socioeconomic activities.

We focus on the number of available hotels in each district before and after the Great East Japan Earthquakes and Tsunami. Especially, we estimate both economic and social damages in three Tohoku prefectures: Iwate(JP-03), Miyagi(JP-04) and

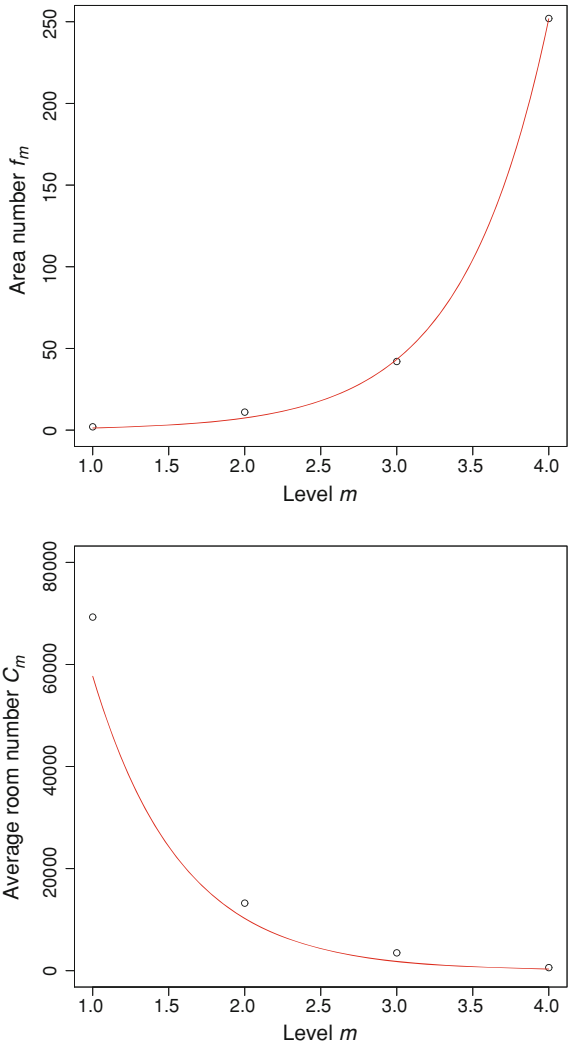


Fig. 7.12 The relation between area number f_m and level m (*top*). The *curve* shows exponential function computed with the least-square method: $f_m = 0.4179e^{1.5849}$. The relation between average room number C_m and level m (*bottom*). The *curve* shows exponential function computed with the OLS regression: $C_m = 324678e^{-1.556}$

Fukushima(JP-07), selecting 21 specfic districts in the three prefectures as shown in Table 7.4 and two periods, which are one before and one after the disaster.

Therefore, we have to estimate the states that were not sampled from these sampled booking data. If we assume that the accommodations included in the data are sampled from uncensored data in a homogeneous way, then the relative frequency of the available accommodation from censored data can approximate the true value, computed from uncensored data. The data on accommodations in this area cover about 31 % of the potential accommodation. Therefore, we have to estimate the uncensored states from these censored booking data.

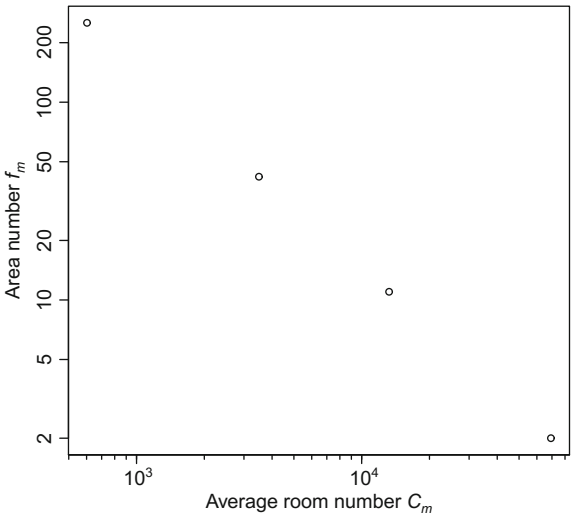


Fig. 7.13 The relation between area number f_m and average room number C_m in a log–log scale. According to Eq. (7.16), the scaling exponent is 1

Table 7.3 List of cities that belong to level 1 and level 2

m	Capacity	Longitude	Latitude	Prefecture
1	88569	35.7340627777778	139.673714722223	Tokyo
	49990	34.5044509594461	135.399519193506	Osaka
2	21068	33.6876391095	130.435037212792	Fukuoka
	19340	42.9922036111203	141.338132222232	Hokkaido
	18261	35.5042238888889	139.620402222222	Kanagawa
	18159	35.1159801925342	136.914723823803	Aichi
	15423	34.9023485760692	135.803759574299	Kyoto
	11911	26.2700305555556	127.729891944444	Okinawa
	10598	38.3092827766947	141.026505832357	Miyagi
	8199	34.3397072222222	132.4663275	Hiroshima
	8147	35.6579983333333	139.876596388889	Tokyo
	7565	35.7407119444444	140.347868055556	Chiba
	6931	33.8997980555556	130.810205833333	Fukuoka

If we assume that accommodation in the data are sampled from uncensored data in a homogeneous way, then a relative frequency of the available accommodations from censored data can approximate the true value that would be computed from uncensored data.

In order to conduct a qualitative study, let $x_i(t, s)$ $(i = 1, \ldots, K; t = 1, \ldots, T; s = 1, \ldots, S)$ be the number of available hotels in district i at day t in period s, where K, T and S represent the number of districts, the number of observations and the number of periods, respectively. Then a relative frequency at district i can be calculated as,

Table 7.4 The ratio of the number of available hotels during the period from 1st to 31st May 2011 to that during the period from 1st to 31st May 2010

Prefecture	District	$q_i(a\|b)$	Complete collapse	Partial collapse	Evacuees
Iwate	Shizukuishi	1.970	0	0	372
	Morioka	1.834	0	4	366
	Appi, Hachimantai, Ninohe	2.250	3	0	0
	Hanamaki, Kitakami, Tohno	1.350	27	364	853
	SanrikuKaigan	0.481	18,098	2,166	12,896
	Oushu, Hiraizumi, Ichinoseki	0.374	83	533	338
Miyagi	Sendai	0.550	21,789	37,522	3,608
	Matsushima, Shiogama	0.345	7,895	12,581	5,115
	Ishinomaki, Kesennuma	0.0	33,661	6,083	23,840
	Naruko, Osaki	1.484	486	1,577	929
	Kurihara, Tome	1.404	224	1,105	1,049
	Shiroishi, Zao	1.608	2,522	1,644	1,612
Fukushima	Fukushima, Nihonmatsu	0.665	168	1,898	1,321
	Soma	0.038	6,279	1,618	1,969
	Urabandai, BandaiKogen	1.134	0	0	2
	Inawashiro, Omotebandai	1.009	10	12	303
	Aizu	1.352	4	27	266
	Minamiaizu	1.768	0	0	14
	Koriyama	0.604	2,596	12,185	2,489
	Shirakawa	1.915	135	1,820	418
	Iwaki, Futaba	0.195	6,550	17,614	2,115

(after and before the Great East Japan Earthquake), the number of both completely destroyed houses and partially destroyed houses, as confirmed at the end of September 2011 and the number of evacuees, as confirmed at 1st May 2011

$$p_i(s) = \frac{\sum_{t=1}^{T} x_i(t,s)}{\sum_{i=1}^{K}\sum_{t=1}^{T} x_i(t,s)}. \tag{7.38}$$

Let us consider a ratio of the relative frequencies after and before a specific event,

$$q_i(a|b) = p_i(a)/p_i(b), \tag{7.39}$$

where $p_i(a)$ and $p_i(b)$ represent the relative frequencies after and before the event, respectively. Obviously, Eq. (7.39) can be rewritten as:

$$q_i(a|b) = \frac{n_i(a)}{n_i(b)} \Big/ \frac{N(a)}{N(b)}, \tag{7.40}$$

where $n_i(s)$ and $N(s)$ are defined as

Table 7.5 The number of evacuees of the Great East Japan Earthquake at three prefectures (Iwate, Miyagi, and Fukushima)

Prefecture	A: public places	B: hotels	C: others	A + B + C
Aomori	0	78	777	855
Iwate	9,039	2,007	14,701	25,747
Miyagi	23,454	2,035	–	25,489
Akita	128	619	909	1,656
Yamagata	305	779	2,366	3,450
Fukushima	6,105	17,874	–	23,979

The data were officially announced by the Japanese Cabinet Office on 3rd June 2011

$$n_i(s) = \sum_{t=1}^{T} x_i(t, s), \quad N(s) = \sum_{i=1}^{K} n_i(s). \tag{7.41}$$

Since $N(a)/N(b)$ is independent of i, $q_i(a|b)$ should be proportional to a ratio of the number of hotels after and before the event.

Table 7.4 shows $q_i(a|b)$, where the term b represents May 2010 (before the disaster), and the term a May 2011 (after the disaster), respectively. Since the value of $q_i(a|b)$ is related to damage to hotels in the district i, $q_i(a|b) < 1$ implies that available hotels decreased after the earthquake at i relative to the total number of hotels. Similarly $q_i(a|b) > 1$ means that they maintained at i.

We may assume that the decrease of $q_i(a|b)$ at district i results from both a decrease of supply and an increase of demand. The decrease of supply is caused in this case by the physical destruction of infrastructure. The increase of demand comes from behaviour of individuals like refugees, workers, volunteers, and civic groups.

The regional dependence of supply can be estimated from the number of destroyed houses in each district. To do so, we calculate the numbers of both completely-destroyed and partially-destroyed houses at each district from the data downloaded from a website of the National Research Institute for Earth Science and Disaster Prevention [18]. The numbers are calculated by summing the number of destroyed houses in the towns or cities included in each district. Table 7.4 shows the numbers of destroyed houses. In this table it is shown that damaged houses were concentrated in the maritime areas of these prefectures.

We can confirm that house damage was serious in Sanrikukaigan, Sendai, Matsushima, Shiogama, Ishinomaki, Kesennuma, Soma, Koriyama, Iwaki, and Futaba. The greatest number of completely-destroyed houses is 33,661 in Ishinomaki and Kesennuma. The second is 21,789 in Sendai. The third is 18,098 in Sanrikukaigan. The greatest number of partially-destroyed houses is 37,522 in Sendai. The second is 17,614 in Iwaki and Futaba. The third is 12,185 in Koriyama.

In fact, in places where the ratio $q_i(a|b)$ is greater than 1, the number of destroyed houses is not significant, as shown in Table 7.4. We confirmed that the ratio $q_i(a|b)$ may measure the degree of damage to economic activity in the travel industry. However, it is not confirmed that there was significant physical damage to houses in

Oushu, Hiraizumi, Ichinoseki, Fukushima, and Nihonmatsu, even having a ratio less than 1. It may be thought that hotels in Oushu, Hiraizumi, and Ichinoseki were used by workers and evacuated victims of the disaster. Decreases of available hotels in Fukushima and Nihonmatsu may be related to accidents in Fukushima Daiich nuclear power plant. The number of victims evacuated from the disaster in each prefecture, according to an official announcement by the Japanese Cabinet Office on 3 June 2011, is shown in Table 7.5. In the case of Fukushima prefecture, 17,874 people were evacuated to hotels at that time. We can see the detail number of evacuee from the web page of the three prefecture [15–17].

7.6 Conclusions

We analysed the data of positions and the number of rooms collected from a Japanese hotel booking site and showed the regional stay capacity and its hierarchical structure.

Firstly, we found that a stay capacity becomes larger as city size is increasing and that a rank size distribution shows power-law relationship. Secondly, we proposed a mathematical method to divide a district into sub-districts with respect to the stay capacities at each district.

It was concluded that the rank size distribution for the number of rooms in the cluster is fitted with a power-law function and that the scaling exponent is dependent on the number of clusters. One of future works is to develop a centroid method regarding that the earth is spherical [2]. This ensures that we can calculate the stay capacities of the areas both more correctly and more globally.

Furthermore, we examined the aftereffect of Great East Japan Earthquake and tsunami turmoil on 11 March, 2011 from the hotel availability estimated from the online data obtained from a hotel booking site. It was found that there was a correlation between the hotel availability and physical damage to the infrastructure.

Acknowledgments The author is thankful to Mr. Kotaro Sasaki and Mr. Daichi Tanaka of RECRUIT Co., Ltd (Jalan) for stimulating discussion.

References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
2. Buss, S.R., Fillmore, J.P.: Spherical averages and applications to spherical splines and interpolation. ACM Trans. Graph. **20**, 95–126 (2001)
3. Cha, S., Mccleary, K.W., Uysal, M.: Travel motivations of Japanese overseas travelers: a factor-cluster segmentation approach. J. Travel Res. **34**, 33–39 (1995)
4. Chen, Y.: The rank-size scaling law and entropy-maximizing principle. Phys. A **391**, 767–778 (2012)
5. Cuccia, T., Rizzo, I.: Tourism seasonality in cultural destinations: empirical evidence from Sicily. Tourism Manage. **32**, 589–595 (2011)

6. Dann, G.M.S.: Anomie, ego-enhancement and tourism. Ann. Tourism Res. **4**, 184–194 (1977)
7. Fischer, M.M., Getis, A. (eds.): Handbook of Applied Spatial Analysis Software Tools, Methods and Applications. Springer, Hidelberg (2011)
8. Freeman, R.E.: Strategic Management: A Stakeholder Approach. Pitman, Boston (1984)
9. Haag, G., Weidlich, W.: A Stochastic Theory of Interregional Migration. Geogr. Anal. **16**, 331–357 (1984)
10. Jalan Web service. URL http://www.jalan.net/jw/jwp0000/jww0001.do
11. Japan Tourism Agency of the Ministry of Land, Infrastructure, Transport and Tourism. URL http://www.mlit.go.jp/kankocho/siryou/toukei/shukuhakutoukei.html
12. Patrick, L.H., Stuart, S., Orkin, E., Carey, G.: Estimating unconstrained hotel demand based on censored booking data. J. Revenue Pricing Manage. **1**, 121–138 (2002)
13. Sato, A.-H.: Patterns of regional travel behavior: an analysis of Japanese hotel reservation data. Int. Rev. Financ. Anal. **23**, 55–65 (2012)
14. Sato, A.-H.: Impact of the Great East Japan Earthquake on Hotel Industry in Pacific Tohoku Prefectures: from spatio-temporal dependence of hotel availability. Prog. Theor. Phys. Suppl. **194**, 165–172 (2012)
15. The data is downloaded from a Web page of Fukushima Prefecture. http://www.pref.fukushima.jp/j/hinanjolist0501.pdf. Accessed 6 Sept 2011 (Japanese)
16. The data is downloaded from a Web page of Iwate Prefecture. http://www.pref.iwate.jp/bousai/taioujoukyou/201105011700hinanbasyo.pdf. Accessed 21 Aug 2011 (Japanese)
17. The data is downloaded from a Web page of Miyagi Prefecture. http://www.pref.miyagi.jp/kikitaisaku/higasinihondaisinsai/pdf/5011900.pdf Accessed 8 Aug 2011 (Japanese)
18. The data is downloaded from a Web page of National Research Institute for Earth Science and Disaster Prevention. URL http://www.j-risq.bosai.go.jp/ndis/. Accessed 31 Aug 2011
19. Tkaczynski, A., Rundle-Thiele, S., Beaumont, N.: Destination segmentation: a recommended two-step approach. J. Travel Res. **49**, 139–152 (2010)
20. Weidlich, W.: Sociodynamics: A Systematic Approach to Mathematical Modelling in the Social Sciences. Taylor and Francis, London (2002)

Chapter 8
Tendency of International Air Travels

Abstract This study considers the relationship between the price of flight tickets and their geodesic distance from the departure airport to the destination. Using the data collected from a Japanese flight booking site, I empirically investigated demand-supply situations from parameter estimates of an Nth order polynomial function of the price in terms of the distance on each observation date. An adequate order of the polynomial function is determined by using two kinds of information criteria (AIC and BIC). It is confirmed that the ticket availability strongly depends on the Japanese calendar date and that the parameter estimates also depend on the calendar date. The parameter estimates may correspond to demand-supply situations of the Japanese air travel market.

8.1 Introduction

How many commercial airports are used in passenger plane? Figure 8.1 shows geographical positions of commercial airports used by scheduled flights within March 2013. 3,388 airports are displayed in our planet. The worldwide air transportation network supports the traffic of over three billion passengers travelling between more than 4,000 airports on more than 50 million flights in a year [13].

Imagine how many connections there are between the airports. It is not so easy to capture all the connections across the globe. However, we may estimate flight tendency from the available number of flight tickets.

Various kinds of items and services can be purchased via e-commerce systems. The emergence of Internet applications has had an unprecedented impact on our lifestyle. Recently, an interest in large-scale data on socioeconomic activities has increased [8]. Utilities and preferences of agents in socioeconomic systems and the availability of items and services at such e-commerce platforms should be studied.

Migration processes have been intensively studied in the context of socioeconomic dynamics, with particular interests in quantitative research. Weidlich and

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_8, © Springer Japan 2014

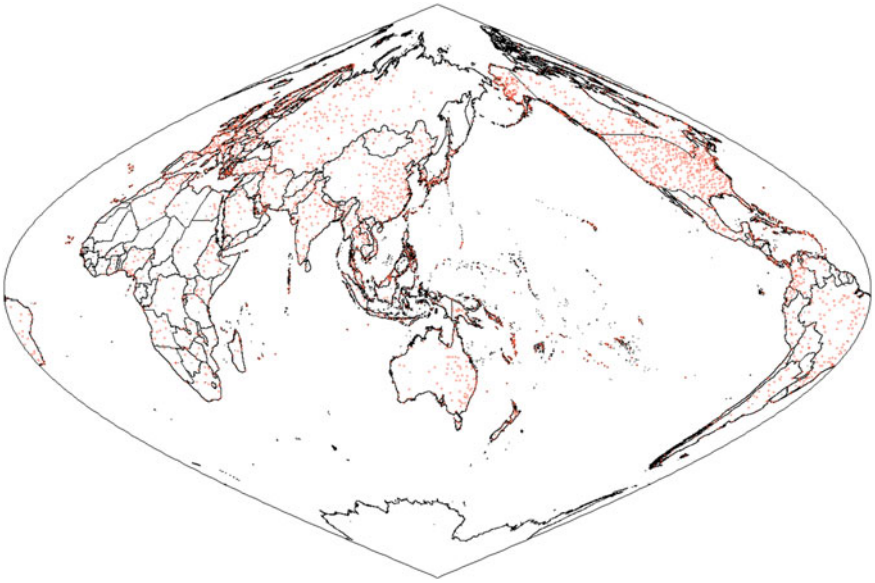


Fig. 8.1 Geographical positions of 3,388 airports used by scheduled flights in March 2013

Haag proposed the Master equation with transition probabilities depending on regional-dependent and time-dependent utility and mobility in order to describe the collective tendency of agent decision in migration choice [7, 12]. Since the motivation to migrate seems to come from both psychological and physical factors, an understanding of the dynamics of migration is expected to lead to knowledge of the inner states of agents and insight to the collective behaviour of agents.

A literature on tourist destination choice pays a great attention to the direct impact of the attributes of the distance to the destination and prices of the destination [9]. There are various approaches to defining a tourist destination. One focuses on destination type, such as regional or national natural parks, scenic or historic sites, hot-spring resort and so forth. Another approach defines choice alternative destinations through aggregation of geographical areas.

Studies of air transportation are necessary in order to understand international tourism management. A relation between demand-supply balance and flight prices is an issue in aviation management. It may be assumed that there are multiple levels of market segments and factors to determine the prices. In principle, geographic, economic, and demographic factors determine the availability of flights (supply) and the potential of passengers (demand). The distance obviously dominates the price of flight tickets. As the distance increases, the price also increases.

Geographic distance is a standard proxy for transport costs under the simple assumption that fees increase monotonically over space. In the case of air transportation, the price of flight tickets and the geodesic distance between the departure and arrival places may be an important issue to be considered in air transportation

management. Usually, the employment cost of crews and the maintenance cost increase as flight distance becomes long. Therefore, there is some relationship between price and distance in aviation.

Sunday studied the parameter effect of prices on American demand from foreign travel and tourism by using regression analysis and panel data [11]. He suggested that high air fares decrease the demand of passengers. Brons et al. examined the price elasticities of passenger demand in air travel. They indicated that long-distance flights generally correspond to higher price elasticities than short-distance flights [4]. In fact, demand-supply situations may influence the flight rates, but it is not obvious that there is relationship among the price of flights, the distance, and the demand-supply situations. Woolley-Meza et al. also investigate the structure and resilience of both the worldwide air-transportation network and the global cargo-ship network [13]. Brockmann and Helbing propose the approach that can identify the spatial origin of spreading processes and be applied to data of the worldwide 2009 H1N1 influenza pandemic and 2003 SARS epidemic based on the international air transpiration network [3].

In this chapter, we focus on the fundamental issue of the relationship between the price of flight tickets and the geodesic distance between the departure and arrival airports. Assuming the Nth order polynomial function of price in terms of the distance, I will estimate the parameters of the relation on each departure date (different demand-supply situations) with the OLS regression and two types of information criteria (AIC and BIC) and examine the relationship between the parameters and the demand-supply situations.

This chapter is organised as follows: In Sect. 8.2, a source of data and data description are explained. In Sect. 8.3, an empirical analysis of the relationship between the price of flight tickets and their geodesic distance is conducted. In Sect. 8.4, the relationship between the price of flight tickets and their geodesic distance is discussed. Section 8.5 is devoted to the concluding remarks.

8.2 Data Description

In this section, I give a brief explanation of a method to collect data on air ticket availability. In this study, I used a Web Application Programming Interface (API) to collect the data. An API is an interface code set that is designed to simplify the development of application programs.

AB-ROAD (http://www.ab-road.net) is a Japanese Internet travel booking site. About 14,000 flight opportunities are available on this site every day. This booking site serves a Web API for both travel agencies and customers. On the one hand, travel agencies can register their flight opportunities on the site via the Internet. On the other hand, consumers can search and book flights that they want to purchase from all the registered flights via the web page. Third parties can even build web services with the data provided by the Web API.

I collected information regarding available flight tickets using the AB-ROAD web service every day and stored it as comma-separated (CSV) files. This data set contains the flight tickets that a person would be able to use to depart from one of the airports in Japan. Each flight also contains the date when I sampled the data, departure date, departure airport, arrival airport, type of class (economy, business, and first classes), name of air carrier, and price (the fuel surcharge and tax are excluded). The data period is from 29 July, 2010 to 14 December, 2011. Due to mechanical reasons, data on several dates is missing (8 November, 2010, 10 April, 2011, from 14 to 25 April, 2011).

Let us denote Δ as the difference between the departure and sample dates. It is inferred that as Δ decreases, the number of flight opportunities decreases. Furthermore, the regional dependence of the number of opportunities on Δ may be related to the supply-demand situation of each destination. We use the data under $\Delta = 28$ days throughout this investigation. In the dataset, there exist about 14,000 kinds of flight opportunities for about 78 airline companies every day.[1]

The total number of available flight opportunities from a city in Japan to a city in a foreign country was counted from the data throughout the entire sampled period. Figure 8.2 shows the total number of flight opportunities per day. From this data, we found three points:

- There exists weekly seasonality for the total number of available flight tickets. The demand of flight tickets is higher on Sundays and Mondays than on other days.
- The number of flight tickets strongly depends on the Japanese calendar. Namely, summer holidays influence the reservation activities of consumers. For example, during Golden week holidays (from 1 to 5 May, 2011) and the holidays in the spring season (around 20 March, 2011), total availability shows steep decreases.
- Since several airline companies update their flight schedule every April and October, the ticket availability drastically drops at that time.

[1] The included airline companies are listed as follows: Jetstar Asia Airways (3K), Cebu Air (5J), Jeju Air (7C), Gill Airways (9C), Jet Airways (9W), American Airline (AA), Air Canada (AC), Mandarin Airlines (AE), Air France (AF), Air India (AI), Aeromexico (AM), Finnair (AY), Alitalia (AZ), British Airways (BA), Eva Air (BR), Air Busan (BX), Air China (CA), China Airlines (CI), Continental Airlines (CO), Cathay Pacific Airways (CX), China Southern Airlines (CZ), Delta Air Lines (DL), Emirates (EK), Etihad Airways (EY), Shanghai Airlines (FM), Garuda Indonesia (GA), Hawaiian Airlines (HA), Hong Kong Airlines (HX), Uzbekistan Airways (HY), Business Air (II), Iran Air (IR), Air Inter (IT), Japan Airlines (JL), JALways (JO), Jetstar Airways (JQ), Korean Air (KE), KLM-Royal Dutch Airlines (KL), Kenya Airways (KQ), Lufthansa German Airlines (LH), Crossair (LX), Air Madagascar (MD), Xiamen Airlines (MF), Malaysia Airline System Berhad (MH), SilkAir (MI), EgyptAir (MS), China Eastern Airlines (MU), All Nippon Airways (NH), Northwest Airlines (NW), Air Macau (NX), Air New Zealand (NZ), MIAT Mongolian Airlines (OM), Austrian Airlines (OS), Asiana Airlines (OZ), Pakistan International Airlines (PK), Philippine Airlines (PR), Air Niugini (PX), Qantas Airways (QF), Qatar Airways (QR), Cargolux (S1), South African Airways (SA), Air Caledonie International (SB), Shandong Airlines (SC), Scandinavian Airlines (SK), Brussels Airlines (SN), Singapore Airlines (SQ), Aeroflot (SU), Thai Airways (TG), Turkish Airlines (TK), Air Tahiti Nui (TN), United Airlines (UA), Air Lanka (UL), Transaero Airlines (UN), Hong Kong Express Airways (UO), Vietnam Airlines (VN), Virgin Atlantic (VS), Vladivostok Air (XF), Arcus Air (ZE) and Shenzhen Air (ZH).

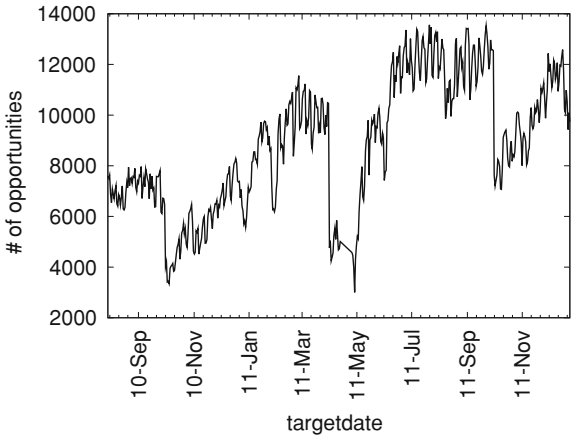


Fig. 8.2 The daily number of flight opportunities from 29 July, 2010 to 14 December, 2011

8.3 Empirical Analysis

Recently, the air transportation network has been studied by several researchers [2, 5, 6, 14, 15]. According to the study by Guimerà and Amaral [6], the world-wide airport network has properties of a small-world network. The degree and betweenness centrality distributions exhibit the power-law decay. In fact, the most connected cities (largest degree) are typically not the most central cities (largest betweenness centrality). Airports with high betweenness tend to play a more important role in keeping networks connected than those with high degree. A passenger can travel from a departure airport to a destination with a short path. Namely, the geodesic distance between departure and arrival airports may give a good approximation of the actual flight distance of passengers.

The geodesic distance is measured by Vincenty's formulae. Let ϕ_s, λ_s, ϕ_f and λ_f be the geographical latitude and longitude of two points s and f, respectively, and $\Delta\lambda = \lambda_s - \lambda_f$. Under the assumption that the earth is a sphere, the distance D between points s and f is approximated as

$$D = r \tan^{-1}\left(\frac{\sqrt{(\cos\phi_s \sin\Delta\lambda)^2 + (\cos\phi_s \sin\phi_f \cos\Delta\lambda)^2}}{\sin\phi_s \sin\phi_f + \cos\phi_s \cos\phi_f \cos\Delta\lambda}\right), \tag{8.1}$$

where r represents Earth's radius ($r = 6371.2\,\text{km}$).

It is possible to analyse the geodesic dependence of ticket prices with this data. Figure 8.3 shows the relation between the price of economy-class flight tickets and the geodesic distance from the departure airport to the destination. The distance of each flight ticket is computed from the geographical latitude and longitude of the departure and arrival airports by using Eq. (8.1). Figure 8.3a represents the relationship of economy-class on 3 August, 2010 (high demand season) and Fig. 8.3b on

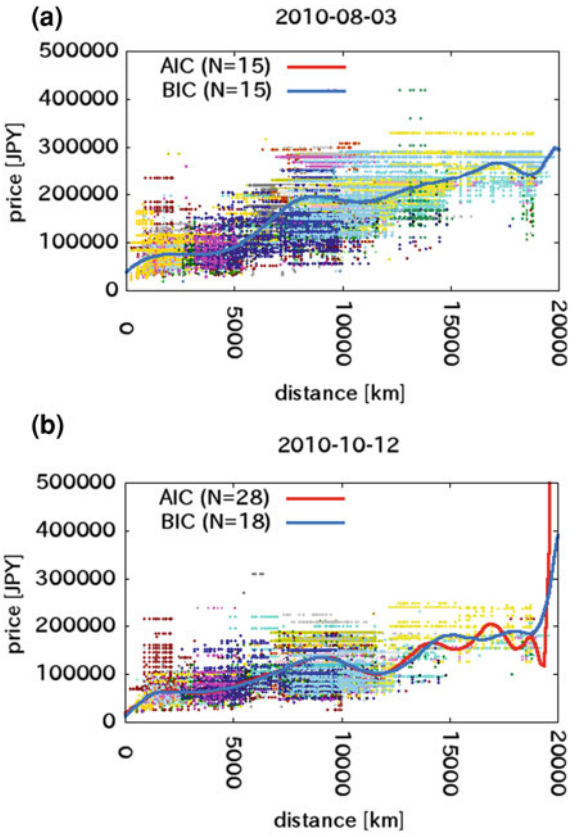


Fig. 8.3 Relationship between the price of economy-class flight opportunities and geodesic distance **a** on 3 August, 2010 and **b** on 12 October, 2010. Each *point* represents the relationship between price and geodesic distance. Each *curve* represents the Nth order polynomial function with parameter estimates by the OLS regression, where the adequate order of the polynomial function is determined by each information criterion (AIC or BIC)

12 October, 2010 (low demand season). Short-distance corresponds to flights to Asian cities (1,000–3,000 km), middle-distance to cities in Europe and North America (8,000–10,000 km), and long-distance to cities in Central and South America (15,000–20,000 km). During high demand season, it is found that various kinds of flights appear for both short-distance and long-distance flights, but, during low demand season, there are few long-distance flights.

Figure 8.4 shows the relationship between the price of business-class flight tickets and the geodesic distance from the departure and arrival airports. Figure 8.4a represents the relationship of business-class on 3 August, 2010 (high demand season) and Figure 8.4b on 12 October, 2010 (low demand season). Since the number of business-class flight tickets for long distance flights (more than 13,000 km) is small, the OLS regression does not seem to work. However, the relations for short distance flights are fitted with the curve.

A demand-supply situation determines price direction. Namely, the excess demand (supply) increases (decreases) prices of goods or services. According to the study by Brons et al. [4], the price elasticities of passenger demand in air travel depend on

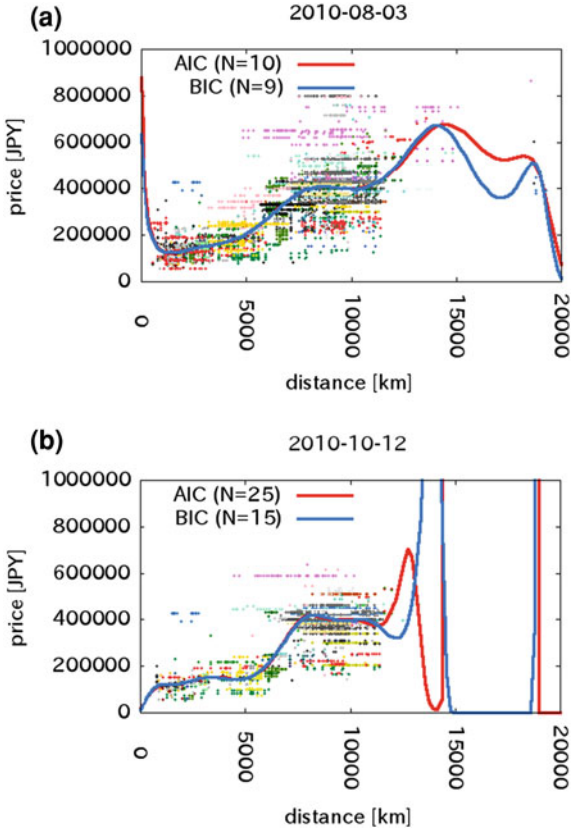


Fig. 8.4 Relationship between the price of business-class flight opportunities and geodesic distance **a** on 3 August, 2010 and **b** on 12 October, 2010. Each *point* represents the relationship between price and geodesic distance. Each *curve* represents the Nth order polynomial function with parameter estimates by the OLS regression, where the adequate order of polynomial is determined by each information criterion (AIC or BIC)

distance. This implies that the price elasticity of demand is a function of distance. Let τ and Q_d denote the price of goods and its quantity demanded, respectively. Suppose that the price elasticity of demand ε_d depends on distance of a flight opportunity D. From the definition of the price elasticity of demand, we may assume

$$\varepsilon_d(D)\frac{\tau}{\mathrm{d}\tau} = \frac{Q_d}{\mathrm{d}Q_d}. \tag{8.2}$$

Therefore, we get

$$\tau(D, Q_d) = cQ^{\frac{1}{\varepsilon_d(D)}} = c\exp\Big(\frac{\ln Q_d}{\varepsilon_d(D)}\Big), \tag{8.3}$$

where c is a positive constant. Equation (8.3) states that price τ is a function in terms of both the quantity of demand Q_d and distance D. The logarithmic form of Eq. (8.3) is described as

$$\ln \tau(D, Q_d) = \ln c + \frac{\ln Q_d}{\varepsilon_d(D)}. \tag{8.4}$$

Expanding Eq. (8.4) in terms of D, we may get

$$\ln \tau(D, Q_d) = \sum_{m=0}^{N} \alpha_m D^m, \tag{8.5}$$

where N denotes the order of the polynomials. The parameters α_m are given by

$$\alpha_m = \begin{cases} \frac{\ln c + \ln Q_d}{\varepsilon_d(0)} & (m = 0) \\ \frac{\ln Q_d}{m!} \frac{\mathrm{d}^m}{\mathrm{d}D^m} \frac{1}{\varepsilon_d(D)} \Big|_{D=0} & (m \neq 0) \end{cases}, \tag{8.6}$$

These parameters depend on the demand-supply situations of both international economics and the seasonal trend of tourism markets. In order to understand such effects on ticket prices, we compute the parameters of the relationship between price and distance for each flight opportunity. The parameters are estimated from the data on each departure date with the OLS regression for Eq. (8.5).

Suppose that there is data on n flight tickets. Let the price of the kth flight ticket and the geodesic distance between departure and arrival places be τ_k and D_k, respectively. Then, a squared error of Eq. (8.5) to the data (τ_k, D_k) $(k = 1, \ldots, n)$ is defined as

$$E(\alpha_0, \ldots, \alpha_N) = \sum_{k=1}^{n} \left(\ln \tau_k - \sum_{m=0}^{N} \alpha_m D_k^m \right)^2. \tag{8.7}$$

Partially differentiating E in terms of parameters α_m $(m = 0, \ldots, N)$, respectively, and setting them into zero, one has

$$\begin{bmatrix} \sum_{k=1}^n D_k^{2N} & \sum_{k=1}^n D_k^{2N-1} & \cdots & \sum_{k=1}^n D_k^N \\ \sum_{k=1}^n D_k^{2N-1} & \sum_{k=1}^n D_k^{2N-2} & \cdots & \sum_{k=1}^n D_k^{N-1} \\ \vdots & & & \vdots \\ \sum_{k=1}^n D_k^N & \sum_{k=1}^n D_k^{N-1} & \cdots & n \end{bmatrix} \begin{bmatrix} \alpha_N \\ \alpha_{N-1} \\ \vdots \\ \alpha_0 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^n D_k^N \ln \tau_k \\ \sum_{k=1}^n D_k^{N-1} \ln \tau_k \\ \vdots \\ \sum_{k=1}^n \ln \tau_k \end{bmatrix}. \tag{8.8}$$

Therefore, I obtain parameter estimates as

$$\begin{bmatrix} \alpha_N \\ \alpha_{N-1} \\ \vdots \\ \alpha_0 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} D_k^{2N} & \sum_{k=1}^{n} D_k^{2N-1} & \cdots & \sum_{k=1}^{n} D_k^{N} \\ \sum_{k=1}^{n} D_k^{2N-1} & \sum_{k=1}^{n} D_k^{2N-2} & \cdots & \sum_{k=1}^{n} D_k^{N-1} \\ \vdots & & & \vdots \\ \sum_{k=1}^{n} D_k^{N} & \sum_{k=1}^{n} D_k^{N-1} & \cdots & n \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^{n} D_k^{N} \ln \tau_k \\ \sum_{k=1}^{n} D_k^{N-1} \ln \tau_k \\ \vdots \\ \sum_{k=1}^{n} \ln \tau_k \end{bmatrix}. \tag{8.9}$$

Assuming Gaussianity of the error term η_k in Eq. (8.5),

$$\ln \tau_k = \sum_{m=0}^{N} \alpha_m D_k^m + \eta_k, \tag{8.10}$$

one obtains the probability density of $\ln \tau$ conditioning on D,

$$p(\ln \tau | D) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln \tau - \sum_{m=0}^{N} \alpha_m D^m)^2}{2\sigma^2} \right], \tag{8.11}$$

where σ^2 represents variance of the error η_k. Then, the log-likelihood function of Eq. (8.11) is defined as

$$l(\alpha_0, \ldots, \alpha_N) = \sum_{k=1}^{n} \ln p(\ln \tau_k | D_k) = -\frac{n}{2}\left\{ \ln(2\pi\sigma^2) + 1 \right\}, \tag{8.12}$$

where the approximation $\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (\ln \tau_k - \sum_{m=0}^{N} \alpha_m D_k^m)^2$ is used during the derivation.

Moreover, Akaike's information criterion (AIC) [1] and Bayesian Information criterion (BIC) [10] are employed in order to determine the number of parameters N.

$$AIC = -2l(\hat{\alpha}_0, \ldots, \hat{\alpha}_N) + 2(N+1), \tag{8.13}$$

$$BIC = -2l(\hat{\alpha}_0, \ldots, \hat{\alpha}_N) + N \ln n. \tag{8.14}$$

The adequate order of the polynomials is selected as N if AIC or BIC takes the minimum value at N. By using this procedure, parameter estimates are computed for each departure date. In order to compare polynomials obtained by minimising AIC with those by BIC, two cases are computed.

Table 8.1 shows the estimated order of the polynomial function, squares error, AIC or BIC, and parameter estimates for $m = 0$ and 1 with economy-class data on 3 August, 2010 and 12 October, 2010. Red and blue curves in Fig. 8.3 show Eq. (8.5) with parameter estimates obtained by using AIC and BIC on these two example dates (high demand date and low demand date). AIC and BIC were computed by using Eqs. (8.13) and (8.14). The curves imply that on the high demand date the price of long-distance flights is clearly higher than on the low demand date. There are peaks at $D = 2{,}500$, 8,000, 14,000, and 16,000 km. On the low demand date, long-distance

Table 8.1 The estimated order of the polynomial function and parameter estimates with the economy-class data on 3 August, 2010 and 12 October, 2010

Date	N	E	AIC	α_0	α_1
03/Aug/2010	15	67.81	4159.73	10.57	8.83×10^{-4}
12/Oct/2010	28	67.87	22831.77	9.85	1.04×10^{-3}
Date	N	E	BIC	α_0	α_1
03/Aug/2010	15	67.81	4307.18	10.57	8.83×10^{-4}
12/Oct/2010	18	67.95	23087.26	9.53	2.16×10^{-3}

Table 8.2 The estimated order of the polynomial function and parameter estimates with the business-class data on 3 August, 2010 and 12 October, 2010

Date	N	E	AIC	α_0	α_1
03/Aug/2010	10	23.86	−2605.88	13.68	-4.37×10^{-3}
12/Oct/2010	25	16.28	−4266.64	9.69	5.60×10^{-3}
Date	N	E	BIC	α_0	α_1
03/Aug/2010	9	23.87	−2529.49	13.36	4.95×10^{-3}
12/Oct/2010	15	16.32	−4135.58	9.55	6.34×10^{-3}

flights (greater than 10,000 km) are less than middle-distance flights (8,000 km). The order of polynomials and parameters for $m = 0$ and 1 obtained by using AIC are slightly different from those by BIC. However, these curves are close to each other until middle distance.

Table 8.2 shows the estimated order of the polynomial function, squares error, AIC or BIC, and parameter estimates for $m = 0$ and 1 with business-class data on 3 August, 2010 and 12 October, 2010. Red and blue curves in Fig. 8.4 show Eq. (8.5) with parameter estimates obtained by using AIC and BIC on these two example dates (high demand date and low demand date). AIC and BIC were computed by using Eqs. (8.13) and (8.14).

Figures 8.5 and 8.6 show parameter estimates on each observation date during the period from 29 July, 2010 to 28 December, 2011. Both the squares error and information criterion took larger value during June to August, 2011 than during the previous and successive periods. This implies that during this period, the relationship between the price and the distance differed from other dates. This is related to the mismatch between the demand of passengers and supply of flights. In fact, in the summer season of 2011, Japanese international air travel tendency decreased in comparison with that of 2010. It is confirmed that α_0 took larger value during the high demand season than during the low demand season. α_0 increased steeply during New Year holidays in January, 2011, spring holidays in May, 2011 and October, 2011. From July to September, 2011 α_0 took smaller value than July to September, 2010. α_1 took smaller value during the high demand season than during the low demand season.

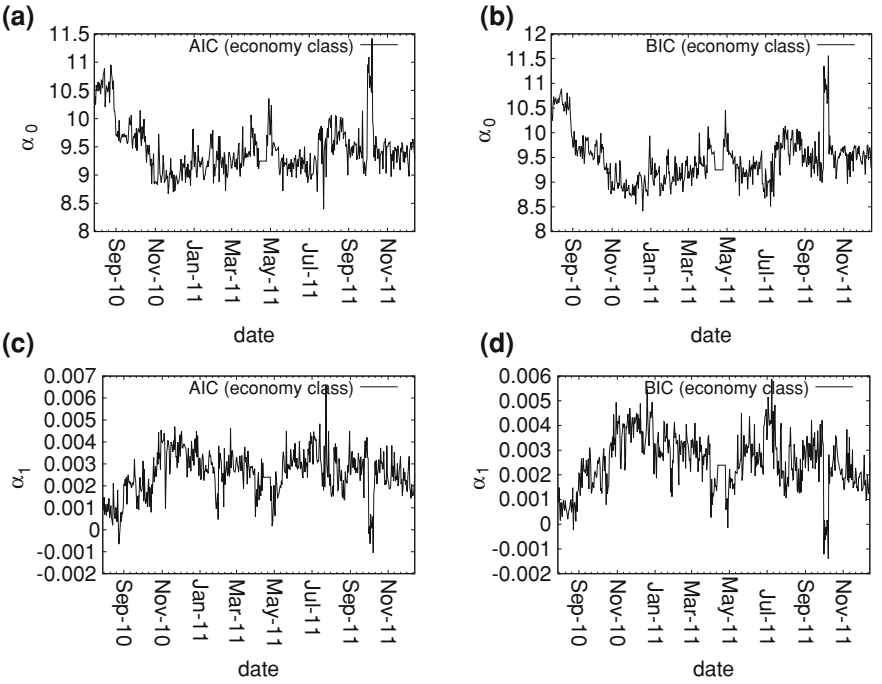


Fig. 8.5 The regression coefficients obtained from the relation between price of economy-class flights and their geodesic distance. **a** α_0 from AIC, **b** α_0 from BIC (economy class), **c** α_1 from AIC and **d** α_1 from BIC by using the OLS regression of Nth order polynomials to the relationship between price and geodesic distance on each departure date during the period of 29 July, 2010 to 14 December, 2011

8.4 Discussion

Since airlines cover their own area of flights, short-distance flights and long-distance flights are managed by different airlines. However, it is found that the price of flights has a tendency to increase as the distance increases as shown in Figs. 8.3 and 8.4. This may imply that the price of flights is not determined by airlines independently, but is adjusted by demand-supply situations. Dominant reasons of this tendency are energy consumption, time duration and competitiveness.

The geodesic distance between the departure and arrival airports is approximately proportional to the energy consumption of the flight. Since recent commercial jet air planes exhaust about 1 kL kerosene fuel to fly 50 km, a passenger exhausts 1 L kerosene fuel to fly 10–15 km. In the case of a 10,000 km distance, a passenger consumes 666 L to 1 kL kerosene fuel. Moreover, the recent commercial jet aeroplanes fly about 800 km/h in velocity. However, the price of air tickets included in the data set exclude the surcharge (fuel charge). Therefore, the price is not related to fuel price directly. The flight price excluding the surcharge is determined by the geodesic dis-

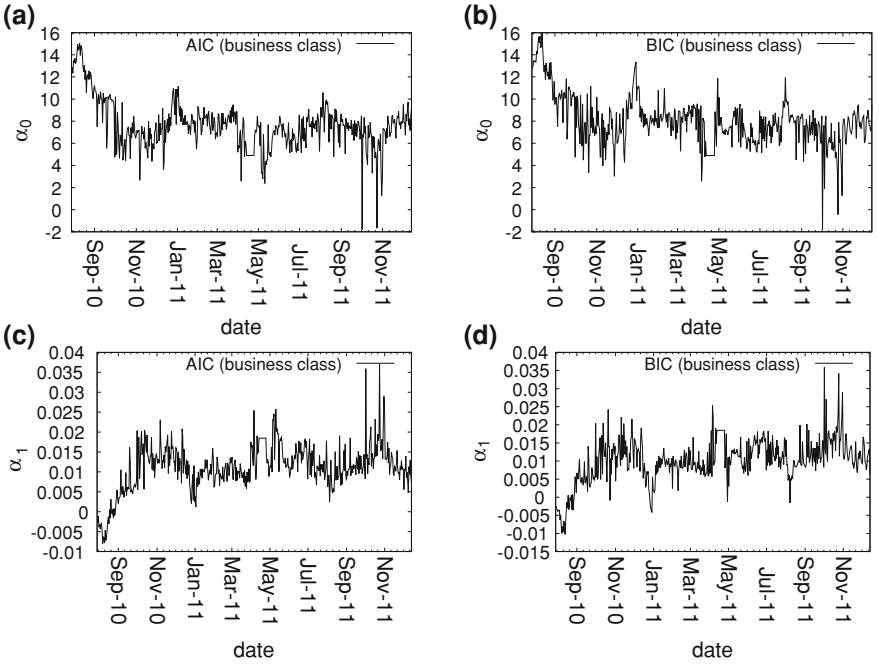


Fig. 8.6 The regression coefficients obtained from the relation between price of business-class flights and their geodesic distance. **a** α_0 from AIC, **b** α_0 from BIC (economy class), **c** α_1 from AIC and **d** α_1 from BIC by using the OLS regression for Nth order polynomials to the relationship between price and geodesic distance on each departure date during the period of 29 July, 2010 to 14 December, 2011

tance, which can be equivalent to time duration of the flight. Suppliers prefer higher price per geodesic distance (time duration) but consumers take lower one. Since there are competitors, one airliner suffers from pressures that prices decline. Of course, it is meaningful to investigate the relationship between surcharge and oil price in order to obtain a better understanding of cost-effectiveness of energy consumptions per capita.

Furthermore, Figs. 8.3 and 8.4 show that the flight prices do not increase linearly to the geodesic distance. Specifically, the flight prices are relatively higher at 3,000, 8,000 and 14,000 km than other distances. These spectra correspond to short distance flights to Asian cities, middle distance flights to European cities, and long distance flights to American cities. There are several cities with higher demand than others. Demand and supply from airports of Japan to these cities are large and their price elasticity is smaller than in other ranges of distance. As a result, there is multimodality in the relation between price of flight ticket and its distance.

Since the flight prices may be determined by both physical factors and demand-supply situations, pricing decisions should be done by air companies independently. However, the competitiveness among airlines plays a role of interaction among them.

Therefore, demand-supply situations can affect the relationship between price of flight tickets and their distance. Its parameter estimates on each observation day may contain the demand-supply situation of Japanese air travel.

As shown in Figs. 8.5 and 8.6, temporal dependence of demand and supply situations is confirmed from the value of α_0 and α_1 for both economy and business classes.

In high (low) demand season the value of α_0 takes a large (small) value. During the summer vacations in 2010 (August to September) α_0 are larger than after the period for both economy and business classes (see Figs. 8.5a, b and 8.6a, b). The values of α_0 in 2011 are less than in 2010. This implies that the demand in the summer holiday season of 2011 is less than in 2010. We further confirm that a peak at New Year Holiday (from the end of December in 2010 to the beginning of January in 2011) for both classes. α_0 exhibits a peak at Golden week Holidays (the beginning of May in 2011) for an economy class specifically (see Fig. 8.5a, b). α_0 during the summer in 2011 does not take lower values than in 2010 for both classes. From this, it is thought that demand of the Japanese air travel market in 2011 was lower than 2011. In August 2011, a peak higher than in 2010 appears for a business class (see Fig. 8.6). Meanwhile, in October 2011, a peak higher than the summer in 2010 appears for an economy class (see Fig. 8.5). However, α_0 for a business class does not show a peak in October 2011 as shown in Fig. 8.6. Large values of α_0 seem to show high demand of flights at this time.

8.5 Conclusion

I collected and analysed data from flight tickets sold on a Japanese flight booking site during the period of 29 July, 2010 to 14 December, 2011. It was found that flight opportunities strongly depend on the Japanese calendar date. It is further confirmed that a relationship exists between the prices of flight tickets (both economy and business classes) and the geodesic distance.

Using the OLS regression for the Nth order polynomials to the data, parameter estimates were computed. The adequate order of equation was selected by using Akaike's information criterion and Bayesian information criterion. The parameters depended on demand-supply situations. High demand seasons (summer vacation, winter vacation, and spring holidays) hold larger α_0 than during the low demand season. It was found that the values of AIC and BIC during the period of June to July, 2011 were larger than those of other observation dates. This is related to the mismatch between demand and supply. This result may help travel agencies understand demand-supply situations of air travel and airlines manage prices of flight tickets and flight schedules from a comprehensive point of view.

Acknowledgments The author is thankful to Prof. Dirk Helbing for his fruitful suggestions and to Ms. Youko Miura (AB-ROAD) for providing useful information on air travel.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (eds.) Proceeding of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Bagler, G.: Analysis of the airport network of India as a complex weighted network. Phys. A **387**, 2972–2980 (2008)
3. Brockmann, D., Helbing, D.: The hidden geometry of complex. Netw. Driven Contagion Phenom. Sci. **342**, 1337–1342 (2013)
4. Brons, M., Pels, E., Nijkamp, P., Rietveld, P.: Price elasticities of demand for passenger air travel: a meta-analysis. J. Air Transp. Manage. **8**, 165–175 (2002)
5. Guida, M., Maria, F.: Topology of the Italian airport network- A scale-free small-world network with a fractal structure? chaos, solitons. Fractals **31**, 527–536 (2007)
6. Guimerà, R., Amaral, L.A.N: Modeling the world-wide airport network. Eur. Phys. J. B **38**, 381–385 (2004)
7. Haag, G., Weidlich, W.: A stochastic theory of interregional migration. Geogr. Anal. **16**, 331–357 (1984)
8. Helbing, D.: Managing Complexity: Insights, Concepts Applications. Springer, Berlin (2008)
9. Nicolau, J.L., Más, F.J.: The influence of distance and prices on the choice of tourist destinations: the moderating role of motivations. Tourism Manage. **27**, 982–996 (2006)
10. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
11. Sunday, A.A.: Foreign travel and tourism prices and demand. Ann. Tourism Res. **5**(2), 268–273 (1978)
12. Weidlich, W.: Sociodynamics-A Systematic Approach to Mathematical Modelling in the Social Sciences. Taylor and Francis, London (2002)
13. Woolley-Meza, O., Thiemann, C., Grady, D., Lee, J.J., Seebens, H., Blasius, B., Brockmann, D.: Complexity in human transportation networks- a comparative analysis of worldwide air transportation and global cargo-ship movements. Eur. Phys. J. B **84**, 589–600 (2011)
14. Zanin, M., Lacasa, L., Cea, M.: Dynamics in scheduled networks. Chaos **19**, 023111 (2009)
15. Zhang, J., Du, W.B., Cao, X.B., Cai, K.Q.: Evolution of Chinese airport network. Phys. A **389**, 3922–3931 (2010)

Chapter 9
Energy Consumption

Abstract The relationship between annual electric power consumption per capita and gross domestic production (GDP) per capita is investigated. In addition, the values of the annual electric power production by four international agencies that report macro data on socioeconomic systems are examined. An increasing tendency of GDP per capita was found in relation to the annual electric power consumption per capita. The results also showed that the data structure, values, and unit depended on the data on annual electrical power consumption in a sample of organisations: the U.S. Energy Information Administration (EIA), International Energy Agency (IEA), OECD Factbook (Economic, Environmental and Social Statistics), and the United Nations (UN) Energy Statistics Yearbook. Further research should establish data standards and an organisation that would oversee to collection, storage, and distribution of data on socioeconomic systems. A distributed energy management system is proposed for the accurate and rigorous collection of data on electrical power consumption.

9.1 Introduction

Sustainability is an important issue throughout the world. The fundamental idea of sustainability was proposed by Buckminster Fuller in his *Operating Manual for Spaceship Earth*, which was first published in 1968 [1]. He proposed that Earth is similar to a spaceship flying through space. He emphasised that the spaceship has a finite amount of resources and the resources that cannot be replenished.

In 1987, the Brundtland Commission proposed a concept of sustainable development. It contains two key ideas: "needs" and "limited resources" in developing countries. The concept of sustainable production next emerged in 1992 at the United Nations Conference on Environment and Development. The conference concluded that especially in industrialised countries, the major cause of the counting deterioration of the global environment is the unsustainable patterns of consumption and production.

Figure 9.1 provides a conceptual illustration of a human society. Energy injection and substantial inflow/outflow are mandatory for maintaining the mechanical and

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_9, © Springer Japan 2014

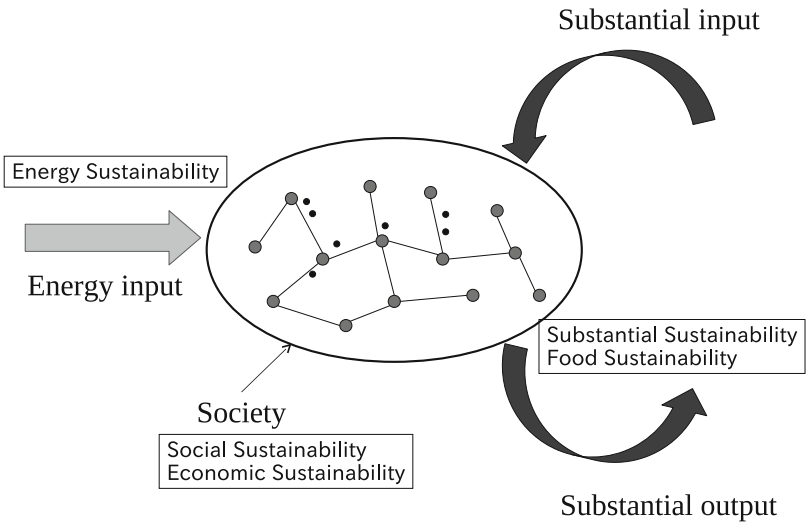


Fig. 9.1 Conceptual illustration of a society consisting of many elements

electrical infrastructure of society. Human resources and social organisations should be resupplied to enable the next generation to maintain our socioeconomic systems. Therefore, the sustainability of our society may be classified into several categories: energy sustainability, substantial sustainability, food sustainability, economic sustainability, social sustainability, and so forth.

Veleva and Ellenbecker proposed a framework and methodology to measure sustainable production [9]. Their framework is based on six main aspects of sustainable production:

- Energy and material use (resources)
- Natural environment (sinks)
- Social justice and community development
- Economic performance
- Workers
- Products

They developed indicators of sustainable production (ISPs). The Lowell Center for Sustainable Production defined sustainable production as [6]:

- Non-polluting
- Conserving of energy and natural resources
- Economically viable
- Safe and healthful for workers, communities, and consumers
- Socially and creatively rewarding for all working people

However, to implement their framework precisely in actual situations, we need more data on human activities. Information communication technology (ICT) is

expected to contribute to constructing a sustainable society for the next generation. In principle, ICT enables us to communicate with one another via computer networks. Large amounts of data on human activities can be transmitted through the computer network and accumulated in a data server. This technology has allowed the emergence of data-centric social sciences at this time. This also has potential for reconstructing our social structure from computerised data.

Our society consists of 7 billion individuals and various types of mechanical and electrical equipment. Each element is located in space and has several properties and states. According to Goodchild [2], every human is able to act as an intelligent sensor: hence, the earth's surface is currently occupied by more or less seven billion sensors. We can extract information from data, construct knowledge from information, and hopefully establish wisdom from several pieces of knowledge. Specifically, researchers in the fields of sociology, economics, informatics, and physics are currently focusing on these frontiers, and they have launched the data-centric social sciences in order to understand the complexity of socioeconomic-technological systems [10]. In order to achieve this outcomes, computer simulation environments, data infrastructure, and high performance computing environment are needed and are expected to yield outcomes in the socioeconomic-technological-environmental sciences.

Measuring the properties and states of social elements yields large amounts of data on socioeconomic activities. The number of elements comprising our society is enormous, and the information generated from our society exceeds the cognitive capacity of an individual. In fact, it is difficult to grasp the state of our social environment. However, it is necessary to understand the state of our society precisely and accurately in order to construct a sustainable community.

In this chapter, we focus on data concerning electrical power consumption as a form of energy consumption. Energy production and consumption are useful quantities in measuring socioeconomic activity. Socioeconomic systems are constructed using mechanical and electronic equipment that is driven by electricity or oil. Therefore, the gross energy consumption in a society is expected to be proportional to its socioeconomic activities. The relationship between annual energy consumption and annual gross domestic product (GDP) has been largely studied in the context of designing efficient energy conservation policies. Using data on gross energy inputs and gross national product (GNP) for the USA, Kraft and Kraft's [4] pioneering study reported causality between GNP and energy consumption. Recently Narayan et al.'s study of Granger causality between electricity consumption and real GDP in 93 countries [8]. They reported that in the six most industrialised nations, increasing electrical power consumption may reduce GDP.

Individual activities in electrical power consumption and economic productivity are strongly correlated. One of the aims of this chapter is to elucidate the relationship between electrical power consumption per capita and GDP per capita. Furthermore, we propose that data management is necessary to understand our social states accurately. Another aim of this chapter is to show the inconsistency in data among organisations that report energy statistics. In order to construct rigorous database

of socioeconomic systems, we need to consider both the rules of data and the roles of organisations.

We also need to consider standards of data generation. I will show a prototype of distributed energy management system that allows us to collect data on both human activity and environments. This energy management system consists of central nodes and sensor nodes, which are designed to behave collectively, based on messages from a cloud server.

This chapter is organised as follows. Section 9.2 describes the relationship between annual electrical power consumption per capita and GDP per capita. Section 9.3 provides an example of the data inconsistency in energy consumption among organisations reporting energy statistics. In Sect. 9.4, I propose the conceptual design of a distributed energy management system. I believe that this would enable us to manage electrical power generation and consumption accurately and rigorously. Section 9.5 is the conclusion to this chapter.

9.2 Relationship Between Energy Consumption and Socioeconomic Activity

Energy production and consumption is deeply related to human activities in our society. This fact can be partially confirmed from the relationship between annual GDP per capita (current USD/person) and annual electrical power consumption per capita (kWh/person).

9.2.1 Relationship for 130 Countries

The graph in Fig. 9.2 contains double logarithmic scatter plots that indicate the annual electric power consumption per capita and GDP per capita in 2009 in 130 typical countries.[1] The data were downloaded from the DataBank of the World Bank (http://data.worldbank.org). The graph shows monotonically increasing tendency of

[1] In this data, annual electrical power consumption per capita of Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Belarus, Belgium, Benin, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei Darussalam, Bulgaria, Cambodia, Cameroon, Canada, Chile, China, Colombia, Congo Dem. Rep., Congo Rep., Costa Rica, Cote d'Ivoire, Croatia, Cyprus, Czech Republic, Denmark, Dominican Republic, Ecuador, Egypt Arab Rep., El Salvador, Eritrea, Estonia, Ethiopia, Finland, France, Gabon, Georgia, Germany, Ghana, Greece, Guatemala, Haiti, Honduras, Hong Kong SAR China, Hungary, Iceland, India, Indonesia, Iran Islamic Rep., Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Korea Rep., Kuwait, Kyrgyz Republic, Latvia, Lebanon, Libya, Lithuania, Luxembourg, Macedonia FYR, Malaysia, Malta, Mexico, Moldova, Mongolia, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Nigeria, Norway, Oman, Pakistan, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russian Federation, Saudi Arabia, Senegal, Serbia, Singapore, Slovak Republic, Slovenia, South Africa, Spain, Sri Lanka, Sudan, Sweden, Switzerland, Syrian Arab Republic, Tajikistan, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela RB, Vietnam, Yemen Rep., Zambia, and Zimbabwe are included.

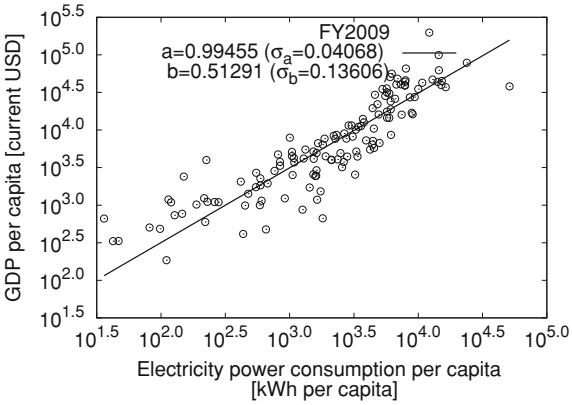


Fig. 9.2 Double logarithmic plots showing annual electric power consumption per capita and GDP per capita. The solid line represents a fitting curve estimated by using the RMA regression

GDP per capita according to the annual electric power consumption per capita. This means that the annual use of electrical power in industrialised countries is greater than in developing countries. Hence, the annual electrical power consumption per capita and the GDP per capita show a positive correlation.

We assume that the GDP per capita y and the annual electrical power consumption per capita x follow an allometric relationship, which is described as the power-law relationship:

$$y = kx^a. \tag{9.1}$$

The parameters a and k are estimated by using a regression for its logarithmic form:

$$\log_{10} y = a \log_{10} x + \log_{10} k. \tag{9.2}$$

Allometric scaling is a symmetrical relationship. Therefore, we estimate parameters a and $\log_{10} k$, not by using the ordinary least squared (OLS) regression, but by the reduced major axis (RMA) regression. Suppose that we have T sets of observations (x_i, y_i). The regression coefficients can be expressed as follows:

$$\hat{a} = \pm\sqrt{\frac{\mathrm{Var}[Y]}{\mathrm{Var}[X]}}, \tag{9.3}$$

$$\hat{b} = \log_{10} \hat{k} = \mathrm{E}[Y] - \hat{a}\mathrm{E}[X], \tag{9.4}$$

and the errors are calculated as

$$\sigma_a = \sqrt{\frac{MSE}{T\mathrm{Var}[X]}}, \tag{9.5}$$

$$\sigma_b = \sqrt{MSE\left(\frac{1}{T} + \frac{\mathrm{E}[X]^2}{T\mathrm{Var}[X]}\right)}, \tag{9.6}$$

Table 9.1 Parameter estimates of the power law relationship between annual electricity consumption per capita (kWh/year/person) and GDP per capita (current USD in 2013 per person)

Year	$\hat{a}$	$\log_{10}\hat{k}$	Error of $\hat{a}$	Error of $\log_{10}\hat{k}$
2000	0.985951758607218	0.300042720350031	0.0482321749097389	0.156301260833071
2001	0.994128210954011	0.259166177818036	0.0469997204333143	0.153298786583552
2002	1.00223642203815	0.239481014780061	0.0461764921803797	0.151081753773536
2003	1.0160249249667	0.241900318713561	0.0458897017838435	0.150819272951516
2004	1.02362467860409	0.264417089625706	0.0454382896190995	0.149844065985442
2005	1.02889543235893	0.282249625648446	0.0447762761005685	0.148349590211332
2006	1.0120719546637	0.380491478362557	0.0427514717779263	0.141849553707711
2007	1.00861045890497	0.450802103823013	0.0414923379717983	0.138055455483469
2008	0.996261007549224	0.547311196947091	0.0407868210738637	0.136310918603779
2009	0.994554527404275	0.5129114058097	0.0406821728454173	0.13606331216819
2010	0.981574478285654	0.577775671218905	0.0405471545958007	0.136193496545949

where the mean square error MSE is computed as

$$MSE = \frac{1}{T-2}\sum_{i=1}^{T}(y_i - \hat{a}x_i - \hat{b})^2 = \Big(\mathrm{Var}[Y] - \hat{a}\mathrm{Cov}[X,Y]\Big)\frac{2T}{T-2}. \quad (9.7)$$

Table 9.1 shows the parameter estimates and their errors for the power-law relationship between GDP per capita and annual electricity consumption per capita for the period from 2000 to 2010. The power law exponent fluctuates around 1.0, which means that the annual electric power consumption per capita is almost proportional to the GDP per capita.

9.2.2 Relationship for 47 Prefectures in Japan

Data on the annual electric power consumption in each prefecture in Japan is available from the Japanese Agency for Natural Resources and Energy of the Ministry of Economy, Trade and Industry.[2] The data on the populations in 47 Japanese prefectures were also downloaded from the homepage of the Statistics Bureau of the Japanese Ministry of International Affairs and Communications.[3] The data on the GDP of each prefecture in Japan is available in the National Accounts of Japan from the Cabinet Office.[4] These data were downloaded from these official sites. Table 9.2 shows the macroeconomic statistics of 47 prefectures in Japan.

Figure 9.3 shows the annual power consumption per capita in each prefecture. The annual electric power consumption per capita and the GDP per capita are computed

[2] Japanese Agency for Natural Resource and Energy of Ministry of Economy (http://www.enecho.meti.go.jp).

[3] Statistics Bureau of Ministry of International Affairs and Communications (http://www.stat.go.jp).

[4] Cabinet Office in Japan (http://www.esri.cao.go.jp/en/sna/memu.html).

Table 9.2 Macroeconomic statistics of 47 prefectures in Japan in 2009

ISO 3166	Name	Oil consumption [ML]	Electricity consumption [GWh]	Heat [TJ]	Population [person]	Area [km^2]	GDP [JPY]
JP-01	Hokkaido	1,812	20,429	10,607	5507,456	83,457	18,052,779
JP-02	Aomori	424	5,448	2,265	1,373,164	9,644	4,416,985
JP-03	Iwate	415	5,401	1,967	1,330,530	15,279	4,254,622
JP-04	Miyagi	600	8,343	2,887	2,347,975	6,862	8,006,517
JP-05	Akita	374	4,819	1,731	1,085,878	11,636	3,697,229
JP-06	Yamagata	373	5,674	1,796	1,168,789	6,652	3,690,958
JP-07	Fukushima	554	8,249	3,502	2,028,752	13,783	7,228,078
JP-08	Ibaraki	426	12,043	6,130	2,968,865	6,096	10,312,413
JP-09	Tochigi	433	12,149	5,172	2,007,014	6,408	7,894,092
JP-10	Gunma	403	12,222	9,907	2,008,170	6,363	7,042,778
JP-11	Saitama	735	26,535	7,301	7,194,957	3,767	20,431,114
JP-12	Chiba	645	20,924	12,707	6,217,119	5,082	19,209,032
JP-13	Tokyo	1,101	59,753	17,686	13,161,751	2,103	85,201,569
JP-14	Kanagawa	806	3,4040	13,930	9,049,500	2,416	29,747,555
JP-15	Niigata	478	12,192	5,363	2,374,922	10,364	8,423,085
JP-16	Toyama	364	7,460	3,134	1,093,365	2,046	4,096,576
JP-17	Ishikawa	349	5,832	2,787	1,170,040	4,186	4,250,003
JP-18	Fukui	256	6,202	4,255	806,470	4,190	3,113,150
JP-19	Yamanashi	174	4,606	1,621	862,772	4,201	2,906,397
JP-20	Nagano	524	11,931	3,527	2,152,736	13,105	7,918,547
JP-21	Gifu	678	11,325	7,486	2,081,147	9,768	6,906,226
JP-22	Shizuoka	895	21,993	17,337	3,765,044	7,329	15,112,757
JP-23	Aichi	1,423	39,777	23,343	7,408,499	5,116	31,891,277
JP-24	Mie	494	12,421	12,895	1,854,742	5,762	7,155,303
JP-25	Shiga	324	9,226	7,447	1,410,272	3,767	5,701,543
JP-26	Kyoto	322	1,1240	5,867	2,636,704	4,613	9,553,851
JP-27	Osaka	783	39,043	13,112	8,862,896	1,898	35,826,529
JP-28	Hyogo	760	24,529	16,080	5,589,177	8,396	17,825,902
JP-29	Nara	143	4,474	2,720	1,399,978	3,691	3,438,173
JP-30	Wakayama	158	4,447	3,079	1,001,261	4,726	3,122,488
JP-31	Tottori	144	2,387	1,062	588,418	3,507	1,888,277
JP-32	Simane	208	3,439	3,024	716,354	6,708	2,333,570
JP-33	Okayama	441	9,164	8,071	1,944,986	7,010	6,928,690
JP-34	Hiroshima	572	1,3941	4,684	2,860,769	8,479	10,815,045
JP-35	Yamaguchi	435	8,735	14,952	1,451,372	6,114	5,476,589
JP-36	Tokushima	165	3,776	3,955	785,873	4,147	2,643,444
JP-37	Kagawa	272	5,237	1,770	995,779	1,862	3,587,627
JP-38	Ehime	307	7,160	7,885	1,430,957	5,678	4,631,968
JP-39	Kochi	145	3,012	289	764,596	7,105	2,140,766
JP-40	Fukuoka	906	20,324	8,446	5,072,804	4,845	17,564,936
JP-41	Saga	231	4,909	1,874	849,709	2,440	2,723,530
JP-42	Nagasaki	336	5,102	2,002	1,426,594	4,105	4,320,061
JP-43	Kumamoto	370	7,540	3,855	1,817,410	7,077	5,366,136
JP-44	Ohita	265	5,871	4,095	1,196,409	5,099	4,044,058
JP-45	Miyazaki	242	5,277	4,021	1,135,120	6,346	3,470,016
JP-46	Kagoshima	474	6,676	1,416	1,706,428	9,044	5,133,170
JP-47	Okinawa	269	4,691	246	1,392,503	2,276	3,721,071

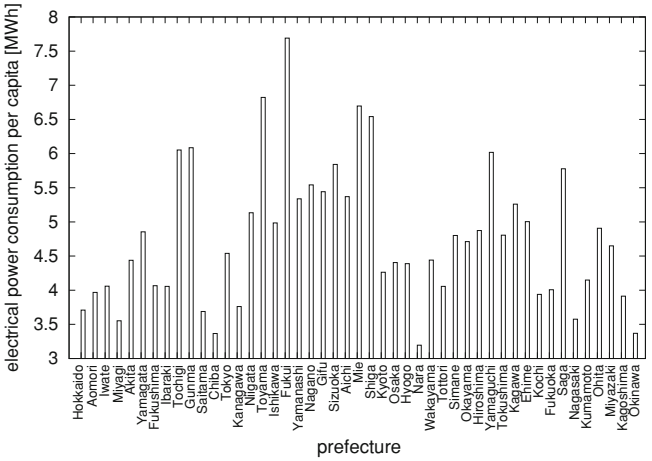


Fig. 9.3 Annual power consumption per capita in 2009 in each prefecture

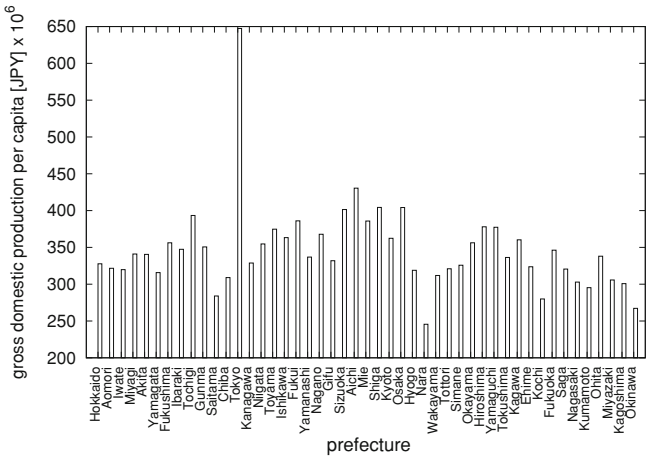


Fig. 9.4 GDP per capita in 2009 in each prefecture

from the values shown in Table 9.2. The annual power consumption per capita in Japan ranges from 3.1 to 7.7 MWh. The highest electricity consumption per capita is in Fukui, at 7.7 MWh, and the lowest is in Nara, at 3.1 MWh. Figure 9.4 shows the GDP per capita in each prefecture. The GDP per capita in Japan ranges from 250 million JPY to 650 million JPY. The highest is in Tokyo, and the lowest is in Nara.

Figure 9.5 shows double logarithmic plots for the annual electric power consumption per capita and GDP per capita of each prefecture in Japan. Throughout Japan, the annual electrical power consumption per capita ranges from 3 to 8 MWh. In fact, Tokyo is an outlier, and the annual GDP per capita in other prefectures shows no correlation with the annual electrical power consumption per capita. These differ-

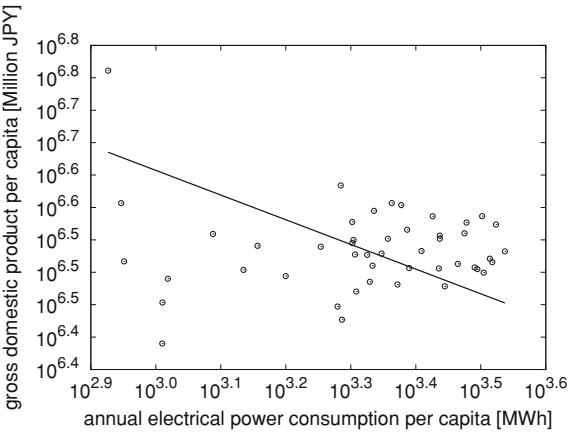


Fig. 9.5 Double logarithmic scatter plots showing the annual electric power consumption per capita and GDP per capita in 2009. A solid curve is computed from a power-law fitting

ences may be because of differences in human behaviour. Specifically, the situation of Tokyo is different from that of other prefectures and is related to the mechanism of incomes of people in Tokyo. The number of company headquarters in Tokyo is larger than in other prefectures and the domestic production in Tokyo is calculated by the production in branches and factories located in other prefectures. However, the slope between the annual electrical power consumption per capita and GDP per capita is less than in the international relationship, as shown in Fig. 9.2. We also obtained $\log_{10} k = 20.4824$ (3.4602) and $a = -2.6261$ (0.5294) with the RMA regression. The power law exponent a in Japan is less than in the whole world. However, the population in Tokyo is an outlier. It shows an increase in GDP per capita of 6,473,422.039 (JPY/year) from only 844.53 (MWh/year). Hence, the homogeneity of Japanese is very high except in the Tokyo population.

9.3 Example of Data Inconsistency

According to the data quality management model, six control dimensions of data quality are proposed:

1. completeness
2. accuracy
3. duplicates
4. consistency
5. integrity
6. conformity

Table 9.3 Electric power production in several typical countries by five international organisations in 2009

United States	*Electrical power production in 2009*
EIA	3,950,331,600,000 [kWh]
UN	4,188,214 [GWh]
OECD	4165.4 [TWh]
World Bank	4,165,394,000,000 [kWh]
Germany	*Electrical power production in 2008*
EIA	594,685,400,000 [kWh]
UN	637,232 [GWh]
OECD	631.2 [TWh]
World Bank	631,211,000,000 [kWh]
Japan	*Electrical power production in 2009*
EIA	984,799,000,000 [kWh]
UN	1,047,919 [GWh]
OECD	1071.3 [TWh]
World Bank	1,040,983,000,000 [kWh]
China	*Electrical power production in 2009*
EIA	3,445,716,000,000 [kWh]
UN	3,714,950 [GWh]
OECD	3 695.9 [TWh]
World Bank	3,695,928,000,000 [kWh]

In the dimension of completeness, key data items are defined in the data structure. In the dimension of accuracy, it is required that the value is consistent with its standard definition. In the dimension of duplicates, only one record exists in the table of key data. In the dimension of consistency, the data in different tables should be consistent with the rule. In the dimension of conformity, the data should follow the standard format.

We found inconsistency in the data on energy statistics of several international organisations. Table 9.3 shows the electrical power production reported by several international organisations, such as the U.S. Energy Information Administration (EIA),[5] the Energy Statistics Yearbook of the United Nations Statistics Division (UN),[6] the OECD Factbook 2011–2012: Economic, Environmental and Social Statistics,[7] and the DataBank of the World Bank.[8]

According to the EIA , the annual production of electricity in the US is estimated at 3,950,331,600,000 kWh. The UN reported that the annual production of electricity

[5] U.S. Energy Information Administration (EIA) (http://www.eia.gov/).

[6] Energy Statistics Yearbook of United Nations Statistics Division (UN) (http://unstats.un.org/unsd/energy/yearbook/default.htm).

[7] OECD Factbook 2011–2012: Economic, Environmental and Social Statistics (http://www.oecd-ilibrary.org/economics/oecd-factbook_18147364).

[8] DataBank of World Bank (http://data.worldbank.org).

in US in 2009 was 4,188,214 GWh. However, the OECD Factbook reported that the annual generation of electricity in the US in 2009 was 4165.4 TWh. The UN Energy Statistics Yearbook reported that the annual generation of electricity in the US in 2009 was 4,188,214 GWh, whereas the World Bank reported it at 4,165,394,000,000 kWh. The same tendency towards inconsistency was confirmed in other countries.

We found that the unit of annual electricity generation is not standardised. The EIA uses kWh, the UN uses GWh, the OECD uses TWh, and the World Bank uses kWh. The values reported by the EIA, UN, and OECD are not the same, but the OECD and the World Bank showed the same values. This means that these statistics are not unique, that is, the values depend on the associations that report data. These associations do not seem to communicate with each other or adjust their reports accordingly. This lack of communication is because of weak international standards regarding the collection and sharing of data on electrical power consumption. Furthermore, no organisation controls or negotiates the data standards. In addition, the updating of data is infrequent because it is delayed for 1–2 years.

In the case of Japan, the Statistics Bureau of the Japanese Ministry of Internal Affairs and Communications collects both micro and macro data on Japan and shares them in a website called E-stat. We suggest that a standard of macro data in socioeconomic systems and several international organisations responsible for socioeconomic data are required.

9.4 Technological Contribution to Energy Management

Recent technologies on smart grids have been intensively developed. Examples are automated meter reading (AMR) and smart meters [3, 5]. ICT may assist the automated matching of electrical power demand and supply. This automated matching system is called a smart grid, which could also be used to measure electrical power generation and consumption in real-time. However, we need to carefully consider balancing consumer privacy with novel applications in the smart grid [7].

We should propose a decentralized distributed energy management system (DDEMS) that does not require the details of data related to the consumer privacy. I propose a concept of DDEMS in Fig. 9.6a. The DDEMS consists of central nodes and sensor nodes. The sensor nodes collect data on energy consumption and the environment (e.g. light intensity, temperature, humidity, and so on). The center nodes accumulate and store data obtained from sensor nodes.

They also control energy balance and communicate through messages from a cloud server. The sensor nodes generate data that transform physical quantities to digital sequences. The communication between the central nodes and the sensor nodes is implemented by using power line communication. The cloud services can control demand and supply and manage the accounting process. DDEMS may contribute to enhancing the usage of data collected by the sensors, as well as the efficiency of the social energy balance.

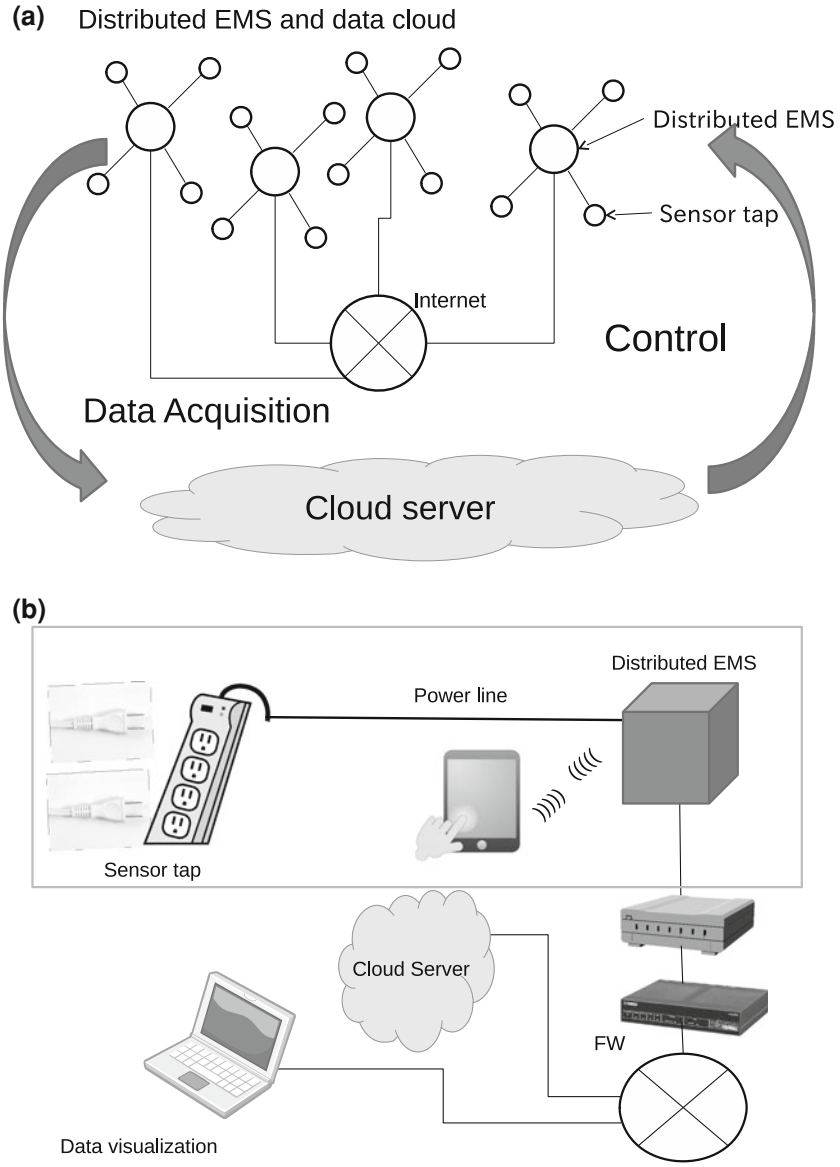


Fig. 9.6 **a** Conceptual illustration of a decentralized distributed energy management system (DDEMS). **b** A schematic illustration of the implementation of DDEMS. The central node (distributed EMS) can be controlled by a user via wireless connections. The sensor node is implemented as a sensor tap, which allows the monitoring and controlling of energy consumption. The central node collects data on energy consumption and the environment in real time through power-line communication

Fig. 9.7 A prototype of sensor node

Figure 9.7 shows prototype of a sensor node. Light intesity and temperature can be recorded. These functioning are implemented by using embidded systems.

9.5 Conclusion

We investigated the relationship between electrical power consumption per capita and GDP per capita in 130 countries using the data reported by World Bank. We found that an electrical power consumption per capita increased as the GDP per increased. The comparison analysis of countries showed a clear scaling relationship. Furthermore, we examined the same relationship in 47 prefectures in Japan. The comparison analysis of 47 prefectures in Japan showed a homogeneity, but less than that found in the 130 countries. This finding may indicate that the relationship between energy consumption and economic activities strongly depends on the life style and social organisation in countries as well as individual.

Moreover, inconsistencies in the data on international electricity production were found in the reports of the EIA, UN, OECD, and World Bank. We suggested the need for data standardisation and the establishment of an organisation that controls the quality and consistency in the international exchange of socioeconomic data.

In this chapter, I proposed a distributed energy management system. This system may contribute not only to managing electrical demand and supply but also to collecting accurate and rigorous data on electrical power generation and consumption. In implementing the proposed central energy management system, we need to carefully consider balancing consumer privacy concerns with novel applications.

Acknowledgments The author expresses his sincere gratitude to Mr. Maito Takagi (DAN Environmental Design Institute Co., Ltd.) and Mr. Tsuyoshi Nagahiro (SSCA) for their stimulating discussions.

References

1. Fuller, R.B.: Operating manual for spaceship earth. Lars Müller Publishers, Baden (2008)
2. Goodchild, M.F.: Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. Int. J. Spat. Data Infrastruct. Res. **2**, 24–32 (2007)
3. IDC Energy Insights: Worldwide quarterly smart meter traker makes its debut; forecasts worldwide growth of 13.0 % to 2015, http://www.idc.com. Accessed 4 March 2012
4. Kraft, J., Kraft, A.: On the relationships between energy and GNP. J. Energ. Dev. **3**, 401–403 (1978)
5. Lennox, J.D.: Automated meter reading: pilot study. Electr. Power Syst. Res. **23**, 47–50 (1992)
6. Lowell Center for Sustainable Production: What is sustainable production? http://www.sustainableproduction.org/abou.what.php. Accessed 7 March 2014
7. McKenna, E., Richardson, I., Thomson, M.: Smart meter data—balancing consumer privacy concerns with legitimate applications. Energ. Policy **41**, 807–814 (2012)
8. Narayan, P.K., Narayan, S., Popp, S.: Does electricity consumption panel granger cause GDP? A new global evidence. Appl. Energ. **87**, 3294–3298 (2010)
9. Veleva, V., Ellenbecker, M.: Indicators of sustainable production: framework and methodology. J. Cleaner Prod. **9**, 519–549 (2001)
10. Weinberger, D.: The machine that would predict the future. Sci. Am. **305**(12), 52–57 (2011)

Part IV
Future Work

Chapter 10
Future Research in Applied Data-Centric Social Sciences

Abstract This chapter addresses future works in applied data-centric social sciences. Rich data on human societies should not only contribute to establishing better understanding of our society but also to developing new services and goods.

10.1 What is Needed to Expand Data-Centric Social Sciences

How many companies are there in the world? How many cars are there in each country? How many buildings are present in each city? How many people are currently flying in aeroplanes? How large an area can be used as residential places? To answer these questions, data is needed on the world. Data on socioeconomic-technological systems are vast, and macroscopic data can be easily accessed. Improving spatial and time resolutions may provide more information on human society, and data infrastructure have recently contributed to enhancing the ability of data.

The human imagination is unlimited. People who lived 100 years ago predicted the shape of recent society. Some of these predictions have been realised, including mobile phones, Internet, air travel, air forces, air combat, air conditioners, television, electricity, bullet train, train network, motorisation and high education. The future of human society has two faces, chance and necessity. Humans also have the ability to design, construct, change and renovate their social systems. Therefore, they can change their world in whichever ways they desire. Specifically, humans need to consider that the scale of human activities may reach the physical limit of their planet. The scope of energy consumption tells us that human activities may reach the physical limit of their planet. Human energy consumption may be influencing climate. Currently, sustainability of society seems to be the most crucial issue in human survival.

To reduce the influence of human activities in environments, human need to change their societal structure. Two possibilities will be addressed:

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1_10,

- to reduce the impact of human activities on their environment, people need to change their societal structures based on Information and Communication Technology.
- to extract or harvest energy from resources not yet in use

Applied data-centric social sciences can contribute to decision-making and finding inefficiency points in various sectors of human society. Humans need to create societal values based on knowledge of data on socioeconomic and technological systems. To do so, they need to have a concept of multi-objective optimisation and real-time data analysis, which is sometimes identified as complex event processing (CEP) in business information systems.

Constructing and deepening links between researchers from fields in data-centric social sciences will provide new insights into how to solve societal problems in the future. Data on human societies are generated from social contexts. Some data are regulated by physical laws, however, other segments of data are generated from social rules (programmes). Rich data are gradually changing many research fields. Applied data-centric social sciences are transdisciplinary or consist of several fields. Below are fields with the potential to take part in applied data-centric social sciences:

- social informatics
- econoinformatics
- computational sociology
- computational economics
- data engineering and computer sciences
- high-frequency finance and econometrics
- tourism informatics
- complexity sciences
- socioeconophysics
- management sciences and marketing
- design technology
- social technology
- statistics
- disaster prevention
- energy management
- healthcare management
- transportation management

Specifically, event data can empower those who have data to change the structure of organisations, implement new technology, predict future contexts and create economical benefits.

10.2 Create Added-Value From Data

It is necessary to create added-value from data. The speedup of processing or updating information seems to be improving. However, this may or may not imply that societies will improve their quality of life directly. For example, high-speed trading seems to

improve society's ability to exchange goods or services. However, trading higher than the human capacity for recognition is meaningless. Too much information also loses meaningfulness since humans cannot be affected by large quantities of information generated at a higher speed than our cognitive capacity. Artificial control of societies without human demand may not create value, but it may be harmful to human life. Humans need to create the lives they want in society. Added-value is connected to what is meaningful in life. Needs should meet seeds of technology.

10.3 Data Synthesis

Data infrastructures are currently being constructed around the world. UNDS and World Bank DataBank are examples of data infrastructures. However, data are still stored separately, and elements must be reconstructed in order to understand what needs to be known. This study is a type of data synthesis where the data are acquired and collected from different sources and different purposes. Their resolution and coverage are dependent on the observing systems employed. From these data sources, both computation and algorithms need to be designed and the data synthesised in order to realise research purposes and to reach goals. This problem is also referred to as data integration or data fusion. The majority of challenging tasks in data synthesis arise from the data to be fused, imperfection and diversity. The number of combinations among data is very large, thus new findings may be discovered in the synthesised data that no one has yet identified.

10.4 Complex Events Processing

CEP has been studied in business intelligence literature and real-time business monitoring. Challenging tasks in CEP are as follows:

- Autonomous data collection with networked sensors
- Automatic data validation with rules and patterns of data
- Automatic data normalisation without human designers
- Automatic detection of events that generate a trigger signal to drive actuators and notify agents without human administration
- Graphic generation based on information, visualisation of data and quantification of affairs
- Data selection based on statistics to measure a degree of outliers from data
- Interlock technology of devices under network disconnection

These technologies not only contribute to create new services and goods, but they allow us to obtain a better understanding of our society based on data accumulated in CEP systems.

Index

A.-H. Sato, *Applied Data-Centric Social Sciences*,
DOI: 10.1007/978-4-431-54974-1, © Springer Japan 2014